A Corpus-based study of repair cues in spontaneous speech

Christine H. Nakatani

Aiken Computation Laboratory, Harvard University, Cambridge, MA 02138

Julia Hirschberg

2D-450, AT&T Bell Laboratories, Murray Hill, NJ 07974

Short Title: Repair cues in spontaneous speech

Received:

The occurrence of disfluencies in fully natural speech poses difficult challenges for spoken language understanding systems. For example, although self-repairs occur in about 10% of spontaneous utterances, they are often unmodeled in speech recognition systems. This is partly due to the fact that little is known about the extent to which cues in the speech signal may facilitate automatic repair processing. In this paper, acoustic and prosodic cues to such repairs are identified, based on an analysis of a corpus taken from the ARPA Air Travel Information System database, and methods are proposed for exploiting these cues for repair detection, especially the task of modeling word fragments, and repair correction. The relative contributions of these speech-based cues, as well as other text-based repair cues, are examined in a statistical model of repair site detection that achieves a precision rate of 91% and recall of 86% on a prosodically labeled corpus of repair utterances. (This paper appears in the Journal of the Acoustical Society of America, 95 (3), March 1994, pp.1603–1616.)

PACS numbers: 43.72Ja,43.70.B,43.70.Bk,43.70.Fq

INTRODUCTION

Studies of large speech corpora have shown that approximately 10% of spontaneous utterances contain disfluencies involving self-correction, or REPAIRS (Hindle, 1983; Shriberg et al., 1992). Blackmer and Mitton (1991) report a rate of one disfluency per 4.6 seconds for radio talk show callers. Yet repairs are often unmodeled in spoken language systems, causing recognition errors such as those shown in Examples (1) and (2). (Recognizer output presented in these examples was generated by the system described in (Lee et al., 1990). The presence of a word fragment in the 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 in Section III.)

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

(2) *Actual string*: Show me all **informa–** information about aircraft type, Lockheed L one zero one one
Recognized string: Show meal **of make** information about aircraft flight Lockhead L one zero one one

In both of these examples, erroneous content words are introduced into the utterance transcription, resulting in uninterpretable recognition output.

Even when all words in a repair utterance are correctly recognized, failure to detect a disfluency can lead to interpretation difficulties during later processing. In Example (3), the string '*twenty two twenty one forty*' must somehow be interpreted as a flight arrival time.

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

Due to an intonational phrase boundary between "two" and "twenty", it is clear to a human listener that the speaker intended the hearer to replace the string "twenty two" with the string "twenty one". The recognition system likewise must choose on some basis among the possible arrival time interpretations, '22:40', '21:40', or '1:40'.

Although undetected and uncorrected disfluencies may lead to serious errors in utterance transcription and interpretation, relatively little attention has been paid to developing methods for automatically detecting and correcting disfluencies for spoken language systems. Robust parsing methods and coarse-grained interpretation strategies may partially buffer a system against these types of errors and interpretation difficulties (Ward, 1991), *inter alia*, but use of these heuristic techniques to process output with recognition mistakes, while of practical use in the short-term, begs the underlying question concerning the acoustic and prosodic nature of this spoken language phenomenon.

Before great energies are spent on solving problems created by deficiencies in speech processing technology, it seems reasonable to inquire whether the technology itself can be enhanced, in this case by direct modeling of repair disfluencies in speech recognition and spoken language understanding systems. In this paper, we contribute to this goal of modeling repair disfluencies by presenting findings on the acoustic and prosodic properties of repairs. We report results from a study of repairs in spoken language system dialogues. Our investigations are guided by past computational and speech analysis work, which we discuss in Section I. Our findings are interpreted within our model of repairs, the REPAIR INTERVAL MODEL (RIM), which we describe in Section II. In Section III, we present our empirical results. In Section IV, we investigate the usefulness of some of our empirical findings in a statistical model of repair detection using CLASSIFICATION AND REGRESSION TREE (CART) techniques, and finally, in Section V, we discuss areas for future work.

I. Previous Work

While self-correction has long been a topic of psycholinguistic study, computational work in this area has been sparse. The methods that have been proposed are, for the most part, text-based — that is, based on the orthographic transcription — and make limited reference to acoustic or prosodic information.

A. Computational Approaches

Early work in computational linguistics treated repairs as one type of ill-formed input. Methods for interpreting repair utterances were developed by extending existing text parsing techniques such as augmented transition networks (ATNs), network-based semantic grammars, case frame grammars, pattern matching and deterministic parsing (Weischedel and Black, 1980; Carbonell and Hayes, 1983; Hindle, 1983; Weischedel and Sondheimer, 1983; Fink and Biermann, 1986). For example, Carbonell and Hayes (1983, p. 128) proposed the following pattern-matching approach to repairing "broken-off and restarted utterances": (a) if two constituents of identical semantic and syntactic type are found when only one is allowed by the grammar, ignore the first one; and (b) recognize explicit corrective phrases (such as "I mean") and appropriately substitute material following the corrective phrase for material preceding the phrase. Constraining the application of both rules was the following meta-rule: "Select the minimal constituent for all substitutions" (Carbonell and Hayes, 1983, p. 128). These rules correctly handle certain cases of repair, but natural language grammars will inevitably allow sequences of the same semantic and syntactic type. Repairs involving these parts of the grammar, such as the cardinal noun phrase repair in Example (3), are left unaddressed. In fact, these kinds of utterances may be ambiguous between repair and non-repair interpretations, such as "**twenty**– two twenty" vs. "twenty two twenty", or "Pick up the **blue**– green ball" vs. "Pick up the blue green ball".

More recently, Shriberg et al. (1992) and Bear et al. (1992) have proposed a two-stage pattern-matching approach to processing repairs. In the first stage, lexical pattern matching rules operating on orthographic transcriptions are used to retrieve candidate repair utterances. Candidate utterances are retrieved by finding an exact repetition of some lexical item or items within a window of n words; a pair of pre-specified adjacent lexical items, such as "a the"; or certain corrective phrases. The candidates are then filtered, using syntactic and semantic information. The first stage of this model, the lexical pattern matcher, was tested on 'nontrivial' repairs, which (Bear et al., 1992) defines as those requiring more editing than deletion of fragments and filled pauses, with the following results reported: 309 of the 406 utterance containing such repairs in their corpus of 10,718 utterances 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%. 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 performing such an experiment.

This two-stage model promotes the important idea that automatic repair processing might be made more robust by integrating knowledge from multiple sources. The lexical pattern matching approach is computationally tractable and provides broad coverage of repair types within a uniform processing framework. In contrast to most earlier work, the effectiveness of the pattern matcher and of several filtering routines was empirically tested on a large corpus of spontaneous speech. However, certain weaknesses of lexical pattern matching in particular and text-based methods in general should be noted.

One such weakness lies in the conceptualization of repair types. It is not clear how to choose among the many possible extensions to an existing set of lexical patterns to increase the coverage of a system, because the principles governing the creation of patterns are not well-specified. Ambiguity in pattern matching also increases the complexity of correction strategies (Bear et al., 1992). A more practical problem for text-based methods in general is their reliance on accurate text transcriptions to identify and correct repair utterances. The assumption that such transcriptions can be produced by existing speech recognizers is optimistic, particularly since current systems rely upon language models and lexicons derived from fluent speech and usually treat disfluencies as noise. One particular challenge that these systems must face is word fragmentation, as exemplified in Examples (1) and (2). Most repair utterances contain word fragments; (Bear et al., 1992) report a rate of 60% (366/607) fragment repairs, while we found 73% (298/382) of repairs in our corpus contained fragments (See Section III.A.2.). However, current recognition systems have no reasonable way of modeling individual word fragments in their lexicons, and thus can output only full words in the lexicon that most closely match the fragment.

Text-based pattern-matching approaches have explored the potential contribution 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*, in addition to particular acoustic-prosodic cues to individual repair patterns, such as suggested in (Bear et al., 1992). One proposal that **does** lend itself to the integration of speech cues into repair detection is that of Hindle (1983), who defines a typology of repairs and associated correction strategies. Hindle identifies the following repair types:

- Full sentence restart: an entire utterance is thrown out and a new utterance is started, e.g. **Is American flight one ninety three** *is dinner served.*
- Constituent level: one syntactic constituent, or part thereof, is replaced by another, e.g. *Show me the cheapest fare* **from Da-** *from Philadelphia to Dallas.*
- Surface level: when identical words are repeated in sequence, the first string of occurrences is thrown out, e.g. *I request uh that you should go to Dallas first uh approximately* **Fri-** *Friday.*

Correction strategies for each repair type are defined in terms of extensions to a deterministic parser. In all cases, the application of a correction routine is triggered by the presence of an hypothesized phonetic "edit signal" marking the point of interruption of fluent speech. This edit signal is described as "a markedly abrupt cut-off of the speech signal" (Hindle, 1983, p.123), following a proposal by Labov (1966). It is treated as a special lexical item in the parser input stream that triggers one of four correction strategies, depending on the parser configuration. Thus, it is the point of detection itself that drives the correction strategy, not simply lexical or syntactic aspects of the repair utterance.

For testing purposes, Hindle employed a corpus of unrestricted spontaneous narratives in which edit signals were orthographically represented and lexical and syntactic category assignments hand-corrected. He found that he could automatically correct 97% of repairs correctly in a corpus of approximately 1500 sentences. Importantly, Hindle's system allows for non-surface-based corrections and sequential application of correction rules (1983, p. 123), which simple pattern-matching approaches cannot readily handle. For example, in (4), a syntactic constituent is replaced by an entirely different one, and in (5), a sequence of overlapping repairs must be corrected.

(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 shows how his copy editing and restart rules, applied deterministically from left to right, are sufficient to handle similarly complex repairs such as Example (6), taken from his narrative corpus.

(6) **I** – **the** – the guys that I **'m** – was telling you about were.

We borrow two main assumptions of Hindle's work in the current study: (a) correction strategies are linguistically rule-governed, and (b) linguistic cues must be available to signal the occurrence of a repair 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 empirical results as evidence supporting this claim, and so we do not concern ourselves in this paper with the study of correction strategies per se. Hindle's results also indicate that, in theory, the edit signal can be computationally exploited for both repair detection and repair correction. However, as Hindle notes, an acoustic-phonetic investigation of repairs is necessary to test the edit signal hypothesis.

Our investigation of repairs is aimed primarily at determining the extent to which repair processing algorithms can rely on the presence of an edit signal in practice. Secondarily, we hope to uncover acoustic and prosodic cues other than the edit signal that may facilitate repair processing, assuming a parsing-based framework such as that outlined by Hindle.

B. Acoustic-Prosodic and Perceptual Studies

Acoustic and prosodic features of repairs have been investigated by psycholinguists, linguists, and other speech scientists. In this section we present an overview of some of these studies. Many of them will be discussed in more detail in later sections.

Early studies of repairs, such as (Nooteboom, 1980; Laver, 1980; Levelt, 1983), examined the phenomenon in the context of proposals for overall cognitive architectures. This work, like much psycholinguistic work on repairs, sought to identify stages of human language production by looking at instances of breakdown and recovery due to self-monitoring. Blackmer and Mitton (1991) measured the timing of repairs in recordings of a Canadian call-in radio program. They concluded that the replanning of speech may occur before the point of self-interruption, in contrast to Levelt (1989) and Laver (1980), who held that replanning commences after fluent speech has been interrupted. Although Blackmer and Mitton observed that many of their repairs involved aberrant phonemes and truncations, no systematic study of these phenomena was made.

Levelt and Cutler (1983) examined the intonational marking of repairs in a corpus of elicited task-oriented speech. They hypothesized that the tendency to mark repairs by accenting the correcting material varies according to semantico-pragmatic properties of the repair utterance. In their corpus, they found that repairs of erroneous information (ERROR REPAIRS) tended to be marked by increased intonational prominence on the correcting information, while other kinds of repairs, such as additions to descriptions (which they labeled APPROPRIATENESS REPAIRS), generally were not.

Bear et al. (1992) noted acoustic differences between true repairs and false positives for repairs that matched two lexical patterns. For repetitions of single words (matching pattern $M \mid M$), true repairs and false positives were reliably distinguished based on the pausal duration within the matched material; for insertions of single words (matching pattern $M \mid XM$), true repairs were distinguished on the basis of changes in FUNDAMENTAL FREQUENCY (f0) and pausal duration. The lack of intonational prominence for repairs matching the first pattern is consistent with Levelt and Cutler's claim, since lexical repetitions are not error repairs. Bear et al. also noted that glottalization may occur at the point of interruption, especially on vowel-final fragments.

O'Shaughnessy (1992) described repairs in a sample of utterances selected from the ARPA Airline Travel and Information System (ATIS) corpus and also investigated pausal duration and intonational prominence as potential correlates of repairs. He found that pausal duration ranged from 100 to 400 milliseconds for 85% of the repairs in his sample. He further reported that repeated words were either uttered with little prosodic change (consistent with previous findings on repetitions) or were shortened by up to 50% of their expected duration, and that substituted or inserted words which added new semantic content to the discourse were intonationally marked in terms of lengthening and higher fundamental frequency, consistent with (Levelt and Cutler, 1983).

Howell and Young (1991) analyzed the pausal and intonational characteristics of repairs in a corpus of conversations between two or more speakers. They identified pause at the interruption site and an increase in intonational prominence at the start of an altered or corrected word as common repair features. These prosodic markings occurred less frequently in repairs involving lexical repetition. The effect of these cues on human processing was tested in a series of experiments in which subjects listened to synthesized stimuli with these features systematically varied. In one task, subjects were asked to judge the comprehensibility of the synthesized speech. In another task, they were asked to produce the corrected version of the synthesized repair utterance. Results showed that pauses and increased intonational prominence helped listeners process repairs, although the facilitative effect of pauses alone was stronger than that of marked accenting alone. Also, the facilitative effects of both cues were stronger for repairs involving lexical alterations than for those involving only lexical repetitions, which is consistent with the results of the initial descriptive study.

Finally, Lickley and colleagues (1991; 1992) carried out perceptual studies on human repair detection using naturally occurring stimuli. Results showed that subjects generally were able to detect a repair before lexical access of the first word in the continuation of fluent speech.

The studies mentioned above describe acoustic and prosodic repair phenomena, and test these findings in perception experiments or on small corpora. Some also investigate how certain phenomena are correlated with repair types. However, researchers have only begun to address issues concerning the modeling of repair phenomena in speech recognition systems and the design of algorithms and methods for automatic repair detection and correction in spoken language systems. Our investigation of repairs addresses these areas of repair modeling and of algorithm development by (a) identifying robust cues to repair that do not rely on sophisticated understanding of the context or the classification of the repair and thus may be detected "on-line" during speech recognition, and (b) exploring empirical methods, namely statistical prediction models, for integrating various cues to achieve automatic repair detection.

II. The Repair Interval Model

To provide a framework for our investigation of acoustic-prosodic cues to repair detection, we earlier proposed a model of repairs, the REPAIR INTERVAL MODEL (RIM) (Nakatani and Hirschberg, 1993a; Nakatani and Hirschberg, 1993b; Hirschberg and Nakatani, 1993). The RIM model divides the repair event into three temporal intervals and identifies time points within those intervals that are computationally critical. A full repair comprises three continguous intervals, the REPARANDUM INTERVAL, the DISFLUENCY INTERVAL, and the REPAIR INTERVAL. Following previous researchers, 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 ('e.g., *uh*', '*um*'), or CUE PHRASES (e.g., '*oops*' or '*I mean*'), which indicate the speaker's recognition of his or her performance error. The REPAIR INTERVAL corresponds to 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 (7), 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.

(7) 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.

As noted in Section I.A, Labov (1966) and Hindle (1983) hypothesized that an "edit signal" occurs at a particular disfluent point within repair utterances, a point in our RIM model which we have labeled the IS. However, our findings and recent psycholinguistic experiments (Lickley et al., 1991) suggest that this proposal may be too limited. So, in this work, we extend Labov's and Hindle's notion of the 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 other prosodic cues which occur in the vicinity of the disfluency interval. The RIM model thus incorporates 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.

As noted in Section I.A, previous acoustic-prosodic and perceptual studies have identified various prosodic cues to repair, such as intonational prominence and pausing, that do not necessarily occur at precisely the IS. Guided by these past findings, we also examine in this study potential cues to repair that may occur during the material to be repaired or the repairing material itself. The RIM model thus serves to focus our attention on timepoints and intervals whose usefulness for automatic methods of repair detection and correction has been established by previous computational or psycholinguistic research.

III. Acoustic-Prosodic Characteristics of Repairs

The corpus for our studies consisted of 6414 utterances from the ARPA Airline Travel and Information System (ATIS) database (MADCOW, 1992) collected at AT&T, BBN, CMU, SRI, and TI; these appear to be a subset of the corpora used by (Shriberg et al., 1992) and (Bear et al., 1992). Of the total corpus of 6414 utterances produced by 122 speakers, 346 (5.4%) utterances contained at least one repair; in our pilot study of the SRI and TI utterances only, we found that repairs occurred in 9.1% of the utterances (Nakatani and Hirschberg, 1993a), a rate which is probably more accurate than the 5.4% we find in our current corpus, since repairs for the pilot study were identified from more accurate and detailed transcriptions than were available for the current corpus. We define repair for our purposes as the self-correction of one or more phonemes (up to and including sequences of words) in an utterance. We thus count as repairs utterances in which a speaker repeats a word or partial word, as well as utterances in which filled pauses or cue words occur immediately after the self-interruption. Utterances in which filled pauses or cue words occur without any self-correction were not classified as repairs in this study.

We developed our initial hypotheses from a pilot study of 146 repairs in the SRI and TI databases (Nakatani and Hirschberg, 1993a). These hypotheses were tested on the three additional ATIS databases

(AT&T, BBN, CMU). Orthographic transcriptions of all of the utterances were prepared by ARPA contractors according to standardized ATIS conventions. The speech we examined was labeled at Bell Laboratories for word boundaries and for intonational prominences and phrasing following Pierrehumbert's description of English intonation (Pierrehumbert, 1980). (Pierrehumbert's system distinguishes two levels of prosodic phrasing, the INTONATIONAL PHRASE and the INTERMEDIATE PHRASE; intonational phrases are composed of one or more intermediate phrases, plus a high or low BOUNDARY TONE, which controls the pitch at the edge of the phrase. Intermediate phrases are composed of one or more PITCH ACCENTS from an inventory of six accent types, plus a PHRASE ACCENT, again, high or low, which controls the pitch from the last pitch accent to the end of the intermediate phrase.) Also, each of the three RIM intervals, together with prosodic and acoustic events within those intervals were labeled. Speech analysis was done with Entropic Research Laboratory's WAVES software (Talkin, 1989).

A. Empirical Results: The Reparandum Interval

In the RIM model, the reparandum interval contains the material to be corrected or replaced by the contents of the repair 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. No reliable cues were found at the reparandum onset. However, we did find some potentially useful cues to repairs at the reparandum offset.

1. Onset of Reparandum

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. One potential prosodic cue to the location of this site might be a phrase boundary marking the beginning of the reparandum interval. Analysis of the TI set uncovered little support for this hypothesis, since a prosodic phrase boundary occurred at the reparandum onset in less than half (42.9%) of the utterances.

This lack of prosodic cues at the reparandum onset is consistent with psycholinguistic findings. As noted in Section I.B, recent perceptual experiments indicate that humans are not able to detect an oncoming disfluency as early as the onset of the reparandum (Lickley et al., 1991; Lickley and Bard, 1992). Subjects in these experiments were presented with successively longer portions of utterances containing repairs and were asked to evaluate whether the partial utterance was "fluent" or "unfluent" up to the end of the stimulus. Judgments as to whether the utterance would continue in a "fluent" or "unfluent" manner were also collected. Subjects *were* generally able to detect disfluencies before lexical access of the first word in the repair. In a few cases where the pause was obviously long, or a word was clearly cut off, subjects detected disfluencies before the start of the repair interval. It should be noted however that only a small number of the test stimuli contained reparanda ending in word fragments (Lickley et al., 1991). In any case, results clearly show that human listeners cannot reliably predict upcoming disfluencies in the region of the onset of the reparandum.

2. Offset of reparandum

In our corpus, 73.3% (298/382) of all reparanda end in word fragments. This finding is somewhat higher than Shriberg et al. (1992)'s report that 60.2% of repairs in their corpus contained fragments. Levelt (1983) reports a rate of 22% for spontaneous speech elicited in an instruction-giving task involving only humans and no computer systems. Lickley (1993) reports a rate of 36% for a corpus of spontaneous conversations by six speakers. The disparity among different corpora remains to be accounted for. Nevertheless, any correlation between rate of fragmentation and spoken language genre should be of interest to researchers developing cognitive theories of monitoring and repair. 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. In our corpus, it is always the case that when a word is fragmented, it is meant to be replaced by some item in the repair interval. Therefore, the interruption of a word is a sure sign of repair, and so we expect that the ability to distinguish word fragments from non-fragments would be a significant aid for repair detection.

Table 1 shows the broad word class of the speaker's intended word for each fragment, where the intended word was recoverable by the ATIS transcribers.

Table 1 goes here.

Fragmentation at the reparandum offset tended to occur in content words (43%) rather than function words (5%), while 52% of intended words were left untranscribed. Table 2 shows the distribution of fragments in our corpus by length. 91% of fragments were one syllable or less in length. Note that O'Shaughnessy (1992) reports that about three-quarters of the fragments in his sample of the ATIS corpus did not have a completion of the vowel in the first syllable.

Table 2 goes here.

While fragments themselves tend to be very short, it is not the case that the reparanda in which fragments occur are significantly shorter than non-fragment reparanda, where reparandum length is measured in number of words. In Table 3, there is no significant difference between the distributions of reparanda lengths for fragment and non-fragment repairs ($p<.20$, χ=6.03, df=4).

Table 3 goes here.

Over one third of fragment reparanda consist of more than simply the fragment. A simple correction heuristic, such as deleting just the fragment portion of the reparandum, might prove effective in many cases, but will not provide a general solution to fragment repair correction.

Table 4 shows the distribution of initial phonemes for all words in the corpus of 6414 ATIS sentences, and for all fragments, single syllable fragments, and single consonant fragments in repair utterances.

Table 4 goes here.

From Table 4 we see that single consonant fragments that are fricatives occur more than six times as often as those that are 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 are less likely to occur as fragments and fricative-initial words more likely to occur as fragments, relative to the underlying distributions for those classes in the corpus as a whole. Both the overall and repair distributions ($p<.0001$, $\chi = 32.88$, df=4) and the single consonant and single syllable distributions ($p<.0001$, $\chi = 66.27$, df=4) differ significantly.

It is possible that the imbalance of content and function words as transcribed intended words for fragments (Table 1) might be due to the general differences in length between content and function words. However, our finding that 90% of all fragments in our corpus are one syllable or less in length (Table 2) provides evidence against this interpretation, since both content and function words are at least one syllable long in English. It might also be thought that the distribution patterns of the initial phoneme (Table 4) might be explained by the possibility that in our corpus fricatives, for example, occur more often as content words rather than function words. We cannot usefully address this question, however, since for over half of all fragments in our corpus the intended word was not recoverable by ATIS transcribers.

Two additional acoustic-phonetic cues, glottalization and coarticulation, may help to identify reparanda offsets, especially those ending in fragments. Bear et al. (1992) note that irregular glottal pulses sometimes occur at the reparandum offset. Shriberg et al. (1992) report glottalization on 24 of 25 vowel-final fragments. In our corpus, 30.2% of reparanda offsets are marked by what we will term INTERRUPTION GLOTTALIZATION. However, 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. Evidently this acoustic-phonetic cue is not always or exclusively associated with word fragmentation.

Interruption glottalization appears to be acoustically distinct from LARYNGEALIZATION (creaky voice), which often occurs at the end of prosodic phrases; the latter typically extends over several syllables, if not words, at the end of an intonational phrase and is associated with a decrease in energy, and low fundamental frequency (Olive et al., 1993). The glottalization we have observed over fragments in our corpus, on the other hand, generally occurs over only the interrupted syllable, and does not appear to be associated with a sustained decrease in energy and fundamental frequency, in contrast to phrase-final laryngealization.

We suspect that this phenomenon of interruption glottalization is akin to one investigated by Local and Kelly (1986). In their study, Local and Kelly report on the acoustic-phonetic correlates of self-interruption. They identify the phenomenon of HOLDING SILENCES on discourse connectives, which they speculate serve the general communicative function of holding the floor, and thus can be associated as well with cases of repair. They characterize these silences in spontaneous speech as

> initiated by glottal closure and terminated by glottal release with closed glottis being *maintained* during the intervening period. Now this kind of 'closure piece' [...] appear[s] to correlate with holding of turns and the projection that there will be further talk by the same speaker (Local and Kelly, 1986, p. 192).

Also, they report no noticeable decrease in rate or amplitude for holding silences. These properties describe occurrences of conjunctions in their corpus, such as *well, but, so, uh*, although they speculate that the phonetic features of holding silences might be generally available as means for the speaker to locally indicate that he or she intends to continue speaking. Interestingly, for certain cases of repairs and reactions to incursive talk by a conversational partner, Local and Kelly report that creaky voice, or irregular glottal pulses, accompany the glottal closure, again without diminuation of tempo or loudness. The similarities between holding silences and repairs exhibiting interruption glottalization suggest that these phenomena are linked by more general principles governing the mechanisms of spoken language interaction.

One other acoustic-phonetic feature which sometimes characterizes words or word fragments at the end of the reparandum interval is the presence of coarticulatory gestures preceding silence. Sonorant endings of both fragments and non-fragments in our corpus sometimes exhibit coarticulatory effects of an acoustically unrealized subsequent phoneme. A related feature is the lack of phrase-final lengthening effects on the last few segments in the reparandum for many cases of repairs. More generally, both of these features are cues to disfluency in the rhythmic structure of pre-pausal segments. These effects might be used to distinguish the offsets of reparanda from fluent phrase offsets. Acoustic models might directly encode information that would distinguish fragment-final phonemes from fluent phrase-final phonemes. For fricatives immediately preceding silence, for example, one might compare duration, energy, and spectral characteristics. For vowels preceding silence, the presence of certain coarticulatory patterns (e.g. stop closure, velar pinch) might positively identify a vowel as fragment-final, since fluent phrase-final vowels followed by silence show no such effects of coarticulation.

To summarize, in our corpus, most reparanda offsets end in word fragments. Transcribers often cannot recover the intended word from repair fragments in our corpus, but the majority of recovered intended words are content words. Fragments are rarely more than one syllable long, exhibit different distributions of initial phoneme class depending on their length, are sometimes glottalized, and sometimes exhibit coarticulatory effects of acoustically missing subsequent phonemes. Procedures for fragment detection might make use of initial phoneme distributions, in combination with information on fragment length and acoustic-phonetic events at the IS. Inquiry into further acoustic-phonetic properties and 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 segments at the reparandum offset.

B. Empirical Results: The Disfluency Interval

In the RIM model, the DI includes all cue phrases and all filled and unfilled pauses from the offset of the reparandum to the onset of the repair. The literature contains a number of findings concerning these phenomena in the DI, such as (Levelt, 1983; Blackmer and Mitton, 1991; Shriberg et al., 1992; O'Shaughnessy, 1992). While our own findings provide little evidence that cue phrases or filled pauses are reliable markers of repairs, we do find the duration of silent pauses to be a reliable characteristic of the DI. In particular, our data support a new hypothesis associating fragment repairs and the duration of pause following the IS.

1. Filled Pauses and Cue Phrases

Filled pauses and cue phrases have been hypothesized as repair cues by Levelt (1983) and by Blackmer and Mitton (1991). In our corpus, such phenomena occur in the DI for only 9.4% (36/382) of repairs. Interestingly, as shown in Table 5, pause fillers and cue phrases occur significantly more often in non-fragment repairs than in fragment repairs ($p<.0001$, $\chi=16.91$, df=1).

Table 5 goes here.

2. Duration of the Disfluency Interval

Duration of pause following the IS also distinguishes between non-fragment and fragment repairs. Table 6 shows the average duration of 'silent DIs' (i.e. those containing no pause fillers or cue words) compared to that of fluent utterance-internal silent pauses (i.e. those which independent labelers had **not** classified as hesitations, repairs, or other disfluencies) for the TI utterances in our corpus.[1]

Table 6 goes here.

Overall, silent DIs are shorter than fluent pauses ($p<.001$, $t=4.65$, df=1530). If we analyze repair utterances based on occurrence of fragments, the DI duration for fragment repairs is significantly shorter than for non-fragment repairs ($p<.001$, $t=3.67$, df=344). The fragment repair DI duration is also significantly shorter than fluent pause intervals ($p<.001$, $t=5.20$, df=1448), while there is no significant difference between non-fragment repairs DIs and fluent phrase boundaries. So, DIs in general appear to be distinct from fluent phrase boundaries. In particular, pausal duration might be exploited to flag potential fragment repairs.

While we do not make specific claims about the higher-level cognitive processes involved in making repairs, we do note that our findings present new facts to be accounted for by current psycholinguistic theories of monitoring and repair. The association of fragment repairs with shorter, usually unfilled, disfluency intervals suggests that, when a speaker interrupts him or herself in mid-word, less time is required to initiate the production of the repairing material than is the case for non-fragment repairs. It has been widely assumed that the replanning process begins no sooner than the point of interruption (Levelt, 1983; Levelt, 1989). If this assumption is to be maintained, then the phenomenon of word fragmentation somehow must be associated with repair types requiring less replanning than other repair types. Alternatively, our duration findings might be interpreted as support for the alternative notion that the duration of the disfluency interval does not exactly reflect the time required to replan. Rather, a theory of monitoring may allow that incremental planning and replanning of speech occur during both silence and speaking, as suggested by Blackmer and Mitton (1991, p. 175).

Finally, we tested a proposal made by O'Shaughnessy (1992) that pausal duration might be used to identify candidate repair sites. Since this proposal was also based on an analysis of a sample from the ATIS corpus, we would expect similar results. O'Shaughnessy suggests that an upper bound of 400 ms on pause duration can be used to identify the disfluency intervals of potential repairs. For his corpus of 115 ATIS repairs, he reports recall of 70% and precision of 65% using this measure. He also proposes that 80 ms represents a lower bound for IS pauses, although he did not test this in his corpus. We tested his proposals for an upper bound only and for both upper and lower bounds on our corpus of 346 repair utterances.

If we try to distinguish repairs from all other potential boundary sites in this data (N=6150 — including non-repair disfluencies and simple word boundaries, as well as fluent phrase boundaries), the 400 ms cutoff proposal identifies 316 of the 390 observed repairs, for a recall of 81%; however, this criterion produces 4860 false positives, for a precision of 6.1%. Precision improves and recall degrades when the lower bound for pauses is added to the prediction — to 53% recall and 34% precision. We conclude that neither proposal yields very reliable results on our corpus. If we eliminate other junctures (i.e., word boundaries and non-repair disfluences), IS sites can be distinguished from fluent pauses using the simple 400 ms cutoff with 81% recall and 34% precision, while the addition of the lower bound degrades performance to 53% recall and 39% precision. In all cases though, the large number of false positives makes the use of pausal duration alone a rather unreliable criterion for identifying repairs. If, in fact, we look at our entire labeled TI corpus, the 400 ms upper bound for IS pausal duration would select fully 58% of all fluent phrase boundaries as potential repair sites.

O'Shaughnessy proposes that output from this detection method be filtered subsequently by searching for identical spectral-time patterns in the speech signal in the immediate areas on either side of the disfluency interval. This spectral-time pattern matching approach can be viewed as approximating the process of lexical pattern matching at the signal level. Whether spectral-time pattern matching can aid repair detection remains to be seen, but we believe this proposal merits further examination. Our results from statistical modeling of repair detection, discussed in Section IV, support the combination of pattern-matching and pause duration information.

3. Prosodic Marking Across the Disfluency Interval

Several influential studies of acoustic-prosodic repair cues have relied upon lexical, semantic, or pragmatic classification of repair types (Levelt and Cutler, 1983; Levelt, 1983). Levelt and Cutler (1983) claim that repairs of erroneous information (ERROR REPAIRS) are marked by increased intonational prominence on the correcting information, while other categories, such as additions to descriptions (APPROPRIATENESS REPAIRS), generally are not. Results of perceptual studies (Howell and Young, 1991) indicate that humans can indeed make use of marked prominence to correct repair utterances. To examine the possibility that intonational prominence might be used in repair detection, we investigated relative pitch and amplitude across the DI for all repairs in our corpus and compared these to the same measurements for fluent pauses in the ATIS TI corpus.

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. (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.) 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, $p<.01$, $t=2.49$, $df=327$). There was also a small but reliable increase in amplitude across the DI (mean=+1.5 db, $p<.001$, $t=6.07$, $df=327$).

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; of course, this failure to find may have been due to the greater variability in the fluent population. 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 probably too small to help distinguish repairs in general 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. 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 and the distribution's center is only slightly above zero. Note that our loci of measurement do not correspond precisely to Levelt and Cutler's, since we examined the syllables immediately surrounding the disfluency interval.

We would emphasize that the analysis reported above was aimed at the discovery of general cues to repairs. The study by Levelt and Cutler (1983) uncovered only tendencies for markedness. For example, only 53% of their error repairs were judged to be intonationally 'marked'. A study by Howell and Young (1991) showed similarly that intonational 'marking', measured in terms of increased intonational prominence, does not occur consistently in repairs. Howell and Young conducted a careful analysis of relative changes in stress levels for repairs in a spontaneous speech corpus that had been independently annotated for three levels of intonational prominence (primary, secondary, and zero stress). They found that the stress levels for pairs of repeated words or pairs of altered words were the same in 72% of cases. For 24%, the stress on the relevant word in the repair was marked with a higher level of stress, while the stress level on the 'repairing' word was lower in only 4% of cases. It may therefore be a better strategy to use a decrease in prominence to rule out potential repairs instead of using increased prominence to positively identify repairs.

C. Empirical Results: The Repair Interval

Previous studies of disfluency have paid considerable attention to the vicinity of the DI but little to the repair offset. The analyses reported above for the reparandum interval and the disfluency interval concentrated on cues for repair *detection*. Our RIM analysis of the repair interval uncovered one general intonational cue that may be of use for repair *correction*, namely the prosodic phrasing of the repair interval. We found evidence that phrase boundaries at the repair offset can serve to delimit the region over which subsequent correction strategies may operate.

First, we tested the hypothesis that repair interval offsets are marked by the presence of intonational phrase boundaries by examining whether phrase boundaries observed at that offset differed in their occurrence from those observed in fluent speech for the TI corpus as a whole; this corpus had previously been labeled at Bell Laboratories for studies on phrasing by Wang and Hirschberg (1992).

Using Wang and Hirschberg's (1992) phrase prediction procedure, with prediction trained on 478 sentences of read, fluent speech from the ATIS TI read corpus, we estimated whether the phrasing at the repair offset was predictably distinct from this model of fluent phrasing.[2] 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 as identified by ATIS transcribers. Results reported here are for prediction on only the 63 TI repair utterances, since the prediction tree we used had been developed on TI utterances.

We found that in these 63 utterances the repair offset co-occurs with minor or major phrase boundaries for 49% of repairs. For 40% of all repairs, an observed boundary occurs at the repair offset where one is predicted in fluent speech; and for 33% of all repairs, no boundary is observed where none is predicted. For the remaining 27% of repairs, observed phrasing diverges from that predicted by a fluent phrasing model. In 37% of these latter cases, a boundary occurs where none is predicted, and, in the remainder, no boundary occurrs when one is predicted.

We also found more general differences from predicted phrasing over the entire repair interval. 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). In our repair utterances, we observed phrase boundaries at repair offsets which occurred within larger NPs, as in Example (8); actual prosodic boundaries in (8a) and (9a) are indicated by '|', and predicted prosodic boundaries by '||' in (8b) and (9b).

In (8), the boundaries which differ from those predicted for fluent speech surround the modifier '*round-trip*'; it is precisely this modifier — not the entire noun phrase — which is being corrected in this utterance.

(8) a. *Actual phrasing:* Show me all **n–** | round-trip | flights | from Pittsburgh | to Atlanta.
b. *Predicted phrasing:* Show me all || round-trip flights || from Pittsburgh || to Atlanta.

It seems plausible that, by marking off the modifier intonationally, a speaker may signal that operations relating just this phrase to an earlier portion of the utterance can achieve the proper correction of the disfluency.

We also found cases in which intonational phrases observed in repair utterances were much longer than phrases observed in fluent speech, as illustrated in Example (9).

(9) a. *Actual phrasing:* **What airport is it** | **is located** | what is the name of the airport located in San Francisco.

b. *Predicted phrasing:* What is the name || of the airport || located || in San Francisco.

The corresponding fluent version of the repair interval is predicted to contain four intonational phrases. In such cases, the absence of intonational phrase boundaries may serve to identify the entire repair (e.g., '*what is the name of the airport located in San Francisco*') as a substituting unit. Thus, in both these cases, the marked phrasing of the repair interval delimits a meaningful unit for subsequent correction strategies.

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 how such phrasal units corresponded to the repair types in Hindle's typology. We wanted to investigate correspondences between intonational phrasing and syntactic characterization of repair type. We found three major classes of phrasing behaviors. First, for 43% (165/382) of repairs, the repair offset we had initially identified (choosing the strategy of identifying the minimal string-length repair) coincides with a phrase boundary, which can thus be said to mark off the repair interval. Note crucially here that, in labeling repairs which might be viewed as either constituent or lexical, we had originally preferred the shorter lexical analysis by default. Of the remaining 217 repairs, 70% (151/217) have the first phrase boundary after the repair onset at the right edge of a syntactic constituent. It is possible that this set of repairs is more appropriately identified as Hindle's constituent repairs, rather than the lexical repairs we had initially labeled. For the majority of these constituent repairs (77%, 117/151), the repair interval contains a well-formed syntactic constituent (See Table 7). If the repair interval does *not* form a syntactic constituent, it is most often an NP-internal repair (74%, 25/34).

Table 7 goes here.

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 lexical repairs (e.g. Example (8)), phonetic errors, word insertions, and syntactic reformulations.

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, 33% (8/24) of NP-internal constituent repairs occurred within cardinal compounds (e.g. Example (3)), which occur often in the ATIS travel information domain. Nonetheless, the fact that repair offsets in our corpus are marked by intonational phrase boundaries in such a large percentage of cases (83%, 316/382) suggests that this cue may prove quite valuable for repair processing by delimiting the interval over which correction strategies may operate.

D. Summary of RIM Results

Using the RIM framework, we have investigated a number of acoustic-prosodic cues to the identification of repairs in spontaneous speech. Our analysis of repairs in the ATIS corpus indicates that self-interruption may be signalled by a number of different cues, including word fragmentation, glottalization, coarticulatory effects preceding silent pauses, and the duration of the disfluency interval itself. We have identified several features to aid in fragment identification, such as the distributions of fragments by length and by initial phoneme. In addition to these reparandum interval and disfluency interval cues for repair detection, we have examined the phrasing of the repair interval for possible cues for repair correction. We have determined that repair intervals differ from fluent speech in their characteristic prosodic phrasing, and identified several roles prosody appears to play in delimiting the repair interval for correction strategies. Given these results, we next turn to their potential use in repair detection.

IV. Predicting Repairs from Acoustic and Prosodic Cues

Despite the moderate size of our sample, we were interested in exploring the question of how well the characterization of repairs derived from RIM analysis of the ATIS corpus would transfer to a predictive model for locating the IS of self-repairs in that domain. We also wanted to investigate how acoustic cues might be combined with the sort of text-based cues which have previously been used with some success by others to predict repair locations (e.g. (Bear et al., 1992)) to improve the predictive power of such text-based cues. To this end, we examined 350 ATIS repair utterances, including the 346 used for the descriptive study. We used the 148 TI and SRI repair utterances used in the initial descriptive study (Nakatani and Hirschberg, 1993a) as training data; the additional 202 repair utterances (containing 223 repair instances) were used for testing. In our predictions, we attempted to distinguish repair IS from fluent phrase boundaries (collapsing major and minor boundaries), non-repair disfluencies (which had been marked independently of our study — see note 1) and simple word boundaries. We considered every word boundary to be a potential repair site; we also included utterance-final boundaries as data points, to distinguish fluent interruptions of the speech signal from non-fluent and for consistency with the prior labelings of our fluent utterances. Thus, our goal was to locate self repairs by distinguishing their ISs from all potential ISs in our test data.

Since each utterance in our test set did in fact contain at least one such IS, this experiment was not equivalent to locating self repairs in general spontaneous speech; the ratio of IS to non-IS data points is considerably greater in our test set. However, utterances could contain more than one repair, so the task was not simply to locate the most likely repair site within an utterance. Nonetheless, our findings should be seen more as indicative of the relative importance of various predictors of IS location than as a true test of repair site location.

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) the duration of pause between w_i and w_j; (b) the occurrence of one or more word fragments within the $< w_i, w_j >$ interval; (c) the occurrence of a filled pause in the $< w_i, w_j >$ interval; (d) the amplitude (energy) peak within w_i — both absolute and normalized for the utterance; (e) the amplitude of w_i relative to w_{i-1} and to w_j; (f) the absolute and normalized f0 of w_i; (g) the f0 of w_i relative to w_{i-1} and to w_j; (h) and w_i's accent status (accented or deaccented). We also simulated some simple pattern matching strategies, to see how acoustic-prosodic cues might interact with lexical cues in repair identification. To this end, we looked at (i) the distance in number of words of w_i from the beginning and end of the utterance; (j) the total number of words in the utterance; (k) whether w_i or w_{i-1} recurred in the utterance within a window of three words after w_i; (l) a part-of-speech window of four around the potential IS; and (m) whether, in cases where w_i and w_j were function words, they shared the same part-of-speech (e.g. PREP PREP).

We were unable to test other acoustic-prosodic features that we had examined in our descriptive analysis, since features such as glottalization and coarticulatory effects had not been labeled in our database for regions 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. We used uniform measures for prediction, however, for both repair sites and fluent regions.

We trained prediction trees using Classification and Regression Tree (CART) techniques (Breiman et al., 1984) given these features. CART techniques can be used to generate decision trees from sets of continuous and discrete variables by using sets of splitting rules, stopping rules, and prediction rules. These rules affect the internal nodes, subtree height, and terminal nodes, respectively. At each internal node, CART determines which factor should govern the forking of two paths from that node. Furthermore, CART must decide which values of the factor to associate with each path. Ideally, the splitting rules should choose the factor and value split which minimizes the prediction error rate. The splitting rules in the implementation employed for our study (Riley, 1989) approximate optimality by choosing at each node the split which minimizes the prediction error rate on the training data. In this implementation, all these decisions are binary, based upon consideration of each possible binary partition of values of categorical variables and consideration of different cut-points for values of continuous variables. Stopping rules terminate the splitting process at each internal node. To determine the best tree, this implementation uses two sets of stopping rules. The first set is extremely conservative, resulting in an overly large tree, which usually lacks the generality necessary to account for data outside of the training set. To compensate, the second rule set forms a sequence of subtrees. Each tree is grown on a sizable fraction of the training data and tested on the remaining portion. This step is repeated until the tree has been grown and tested on all of the data. The stopping rules thus have access to cross-validated error rates for each subtree. The subtree with the lowest rate then defines the stopping point for each path in the full tree. Trees described below all represent cross-validated data. The prediction rules work in a straightforward manner to add the necessary labels to the terminal nodes. For continuous variables, the rules calculate the mean of the data points classified together at that node. For categorical variables, the rules choose the class that occurs most frequently among the data points. The success of these rules can be measured through estimates of deviation. In this implementation, the deviation for continuous variables is the sum of the squared error for the observations. The deviation for categorical variables is simply the number of misclassified observations.

The best prediction tree trained on our 148 utterance training set was then used to predict IS boundary locations in our test set. This tree is illustrated in Figure V..

Figure V. goes here.

The variables represented in this tree — those that CART found most useful in predicting the training data — include *pause*, the duration of pause between w_i and w_j; *frag*, the presence of one or more word fragments in the DI between w_i and w_j; *filled*, the presence of a filled pause in the DI; *lex*, a repetition of w_i within a window of three words to its right (measured in terms of distance in words of the repetition from w_i); *prevlex*, a repetion of w_{i-1} within a window of three to the right of w_i (also measured by distance from w_i); *dups*, duplication of function words with the same part of speech across the potential disfluency site; *j4f*, the part of speech of w_{j+1}; and *f0*, the peak f0 of w_i. The node labels are *phrase*, a fluent phrase boundary; *is*, the IS of a self-repair; *odisfl*, the site of a non-repair disfluency; and *na*, for simple word boundaries.

When this tree is used to predict the test set, it identifies 192 of the 223 observed repairs, with 19 false positives, representing a recall of 86.1% and precision of 91.2%. Note that all repairs are identified in part by the duration of the interval between w_i and w_j. Fully 106 of the correctly identified ISs were also distinguished by the presence of word fragments in the DI. Others were identified from (a) pause filler and part-of-speech information; (b) lexical matching across the DI; and (c) duplication of part-of-speech across the DI. The utility of combining general acoustic-prosodic constraints with lexical pattern matching techniques as a strategy for repair identification thus appears to gain support from this experiment.

The prediction tree in Figure V. utilizes certain features hypothesized to be critical for repair identification. For example, the presence of a pause appears at the top of the tree as a strong predictor of repairs. In the tree in Figure V., the presence of a pause is a necessary condition for a repair, although it is not a sufficient one, as discussed in Section III.B.2. The presence of a word fragment at the IS, which characterizes the majority of repairs in our corpus, turned out to be the next strongest predictor in the tree. As proposed in the literature, filled pauses may indicate repair, but since fluent phrases and other disfluencies such as hesitations may also be marked by filled pauses, this cue is more productively used in combination with other lexical and prosodic cues as shown in Figure V.. Finally, the tree nodes lower in the tree that test for repetitions of lexical items or part of speech within a small window, seem to capture lexical pattern matching information.

Several features that had been hypothesized to be strong repair cues do not in fact appear in the prediction tree in Figure V.. There are no nodes for repair identification that specify maximum pausal duration during the DI, or that make reference to fundamental frequency or amplitude values, for example. As discussed in Section III.B.2, we found that pausal duration is most useful in distinguishing fragment DIs from non-fragment DIs, but less useful in distinguishing all DIs from fluent pause intervals. We speculate that the absence of such maximal pausal duration cues for the DI in the tree in Figure V. may be related to the fact that the presence of a fragment was directly represented in our CART modeling, enabling direct classification of fragment repairs. Relative change in pitch and amplitude also have been claimed to be significant repair cues in the literature but do not appear in the prediction tree. In Section III.B.3, we conjectured that since these cues are specific to certain repair types, and even then do not occur obligatorily to mark repair types, they would not serve as robust cues to repair.

Although the practical integration of our acoustic-prosodic findings with existing proposals for repair detection and correction remains to be done, our predictive modeling of repairs in the ATIS domain using CART analysis takes a first step in this direction. Larger corpora must be examined, but our results of 86% recall and 91% precision, while preliminary, provide additional evidence that sufficient cues may exist in the vicinity of the DI to identify the majority of repairs in a local manner.

V. Discussion

As we noted in the introduction, one approach to repair processing is to compensate for speech recognition errors by employing robust parsing and interpretation techniques. Many text-based methods embody this approach by assuming hypothesized strings of text as input to repair processing strategies. In contrast, the motivation for this study was to explore the extent to which the speech technologies themselves may be enhanced. To this end, we developed the REPAIR INTERVAL MODEL to provide a general model of the temporal intervals that comprise a repair, and we explored a variety of acoustic and prosodic signals that may be associated with computationally critical regions of these intervals. Several repair cues identified by our analysis also proved useful in statistical prediction models for repair.

We conclude from our empirical investigations and statistical modeling of repairs that different repair phenomena might be handled most aptly by different speech recognition technologies and techniques. Pausal duration cues to repair could be exploited in word-based recognition systems with accurate silence detection capabilities. Also within the word-based recognition paradigm, spectral-time pattern matching of repeated words might be implemented as an approximation of lexical pattern matching at the signal level. This procedure conceivably could proceed before all word identities are hypothesized and could be triggered by the presence of a pause.

One problem that poses difficulties for word-based methods is that of detecting word fragments. A default strategy that has been proposed for fragment recognition has been that all fragments in a corpus be treated as instances of a single generic token. The length distributions and the wide phonetic and phonemic variation of fragments in our corpus suggest that a more fruitful approach might be to recognize fragments bottom-up rather than top-down. Preliminary investigations suggest that phone-based recognizers may be well-suited to such a task. Informal testing of the phone-based recognizer described in (Ljolje and Riley, 1992) on a subset of our corpus indicated that such a recognizer could identify many of our fragment phonemes in the same manner as non-fragment phonemes — even certain fragment-final phonemes that were heavily coarticulated with their preceding phonemes. However, important theoretical questions remain concerning how fragments may be recognized even given accurate phonemic transcriptions. In what manner do fragments violate phonotactic constraints? Can these be exploited in bottom-up prediction of fragment regions? Careful study of the spectral and durational characteristics of abruptly cut off segments are needed to determine whether these reparandum-final phonemes differ significantly enough from fluent phrase-final phonemes to provide direct acoustic-phonetic evidence of repair.

It appears likely that certain repair phenomena ultimately will receive an explanation in terms of their articulatory bases. The phenomenon of interruption glottalization and its related phenomenon of holding silences, together with other cases of partly articulated phonemes at the IS, may best be modeled in terms of the partially completed articulatory gestures involved in their production.

The phenomena realizing the interruption of fluent speech can effect change at multiple levels of language, from the syntactic constituent to the phone. Thus, our models of repair must provide for phoneme fragments as well as sentence and word fragments. Viewed from this perspective, the speech processing of repairs presents itself as a promising area in which to explore the integration of various paradigms of speech processing in a productively focused manner. Our study illuminates some of the ways in which various aspects of this problem can be directly modeled in speech recognition and spoken language understanding systems.

Acknowledgments

We thank John Bear, Mary Beckman, John Coleman, Barbara Grosz, Don Hindle, Chin Hui Lee, Robin Lickley, Andrej Ljolje, Joe Olive, Jan van Santen, Stuart Shieber, Liz Shriberg, and two anonymous reviewers for advice and useful comments. CART analysis employed software written by Daryl Pregibon and Michael Riley at AT&T Bell Laboratories. The first author was partially supported by a National Science Foundation Graduate Research Fellowship.

Notes

[1]Here and below we treat the fluent ATIS TI utterances as a sample corpus of fluent ATIS utterances against which we compare our repair corpus findings. The full ATIS TI corpus, including both repair and non-repair utterances, was independently labeled for a previous study: Labelers for this study,reported in (Wang and Hirschberg, 1992), were told to mark the following disfluencies: REPAIR (self-correction of lexical material), HESITATION ("unnatural" interruption of speech flow without any following correction of lexical material, including all events with some phonetic indicator of disfluency that were not involved in a self-repair, such as audible breath or sharp cut-off), or OTHER DISFLUENCY (material deemed disfluent but not falling into either of the previous categories). See Wang and Hirschberg (1992) for further information on the labeling of this corpus.

[2]Wang and 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% (estimated) accuracy on the ATIS TI read

corpus using only features whose values could be calculated via automatic text analysis. These utterances contained no disfluencies. The confusion matrix for prediction of boundaries vs. null boundaries is presented in Table 8; note that the 91.7% success rate is higher than the CART estimated accuracy of 88.4%, since the confusion matrix is derived from the full CART tree, and thus should be taken as indicative only of performance predicting boundaries vs. null boundaries:

Table 8 goes here.

Read the table as, for example, 62.8% of data points were correctly identified as 'null boundary'.

References

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

Jaime Carbonell and Pat Hayes. 1983. Recovery strategies of parsing extragrammatical language. *American Journal of Computational Linguistics*, 9(3-4):123–146.

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.

Pamela E. Fink and Alan W. Biermann. 1986. The correction of ill-formed input using history-based expectation with applications to speech understanding. *Computational Linguistics*, 12(1):13–36.

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.

Julia Hirschberg and Christine Nakatani. 1993. A speech-first model for repair identification in spoken language systems. In *Proceedings of the Third European Conference on Speech Communication and Technology*, Berlin, September. Eurospeech-93.

P. Howell and K. Young. 1991. The use of prosody in highlighting alterations in repairs from unrestricted speech. *The Quarterly Journal of Experimental Psychology*, 43A(3).

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

John Laver. 1980. Monitoring systems in the neurolinguistic control of speech production. In V. Fromkin, editor, *Errors in Linguistic Performance*, pages 287–305. Academic Press, New York.

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.

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

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

Willem Levelt. 1989. *Speaking: From Intention to Articulation.* MIT Press, Cambridge MA.

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.

R. J. Lickley. 1993. Personal Communication, June.

A. Ljolje and M. D. Riley. 1992. Optimal speech recognition using phone recognition and lexical access. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, October. ICSLP.

John Local and John Kelly. 1986. Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies*, 9:185–204.

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

Christine Nakatani and Julia Hirschberg. 1993b. A speech-first model for repair detection and correction. In *Proceedings of the 31st Annual Meeting*, Columbus. Association for Computational Linguistics.

S. G. Nooteboom. 1980. Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. Fromkin, editor, *Errors in Linguistic Performance*, pages 287–305. Academic Press, New York.

Joseph Olive, Alice Greenwood, and John Coleman. 1993. *The Acoustics of American English Speech: A Dynamic Approach.* Springer-Verlag, New York.

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

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.

Michael D. Riley. 1989. Some applications of tree-based modelling to speech and language. In *Proceedings of the Speech and Natural Language Workshop*, Cape Cod MA, October. DARPA, Morgan Kaufmann.

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.

David Talkin. 1989. Looking at speech. *Speech Technology*, 4:74–77, April-May.

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

W. Ward. 1991. Understanding spontaneous speech: the phoenix system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 365–367. IEEE.

R. M. Weischedel and J. Black. 1980. Responding to potentially unparseable sentences. *American Journal of Computational Linguistics*, 6:97–109.

R. M. Weischedel and N. K. Sondheimer. 1983. Meta-rules as a basis for processing ill-formed input. *American Journal of Computational Linguistics*, 9(3-4):161–177.

Lexical Class	Tokens	%
Content	128	43%
Function	14	5%
Untranscribed	156	52%

Table 1: Lexical Class of Word Fragments at Reparandum Offset (N=298)

Syllables	Tokens	%
0	119	40%
1	153	51%
2	25	8%
3	1	0.3%

Table 2: Length of Reparandum Offset Word Fragments (N=298)

Length	Fragment Repairs		Non-fragment Repairs	
in words	N (280)	%	N (102)	%
1	183	65%	53	52%
2	64	23%	33	32%
3	18	6%	9	9%
4	6	2%	2	2%
5 or more	9	3%	5	5%

Table 3: Length of Reparandum Interval for Fragment and Non-fragment Repairs

Class of Initial Phoneme	% of All Words	% of All Fragments	% of One Syllable Fragments	% of One Consonant Fragments
stop	23%	23%	29%	12%
vowel	25%	13%	20%	0%
fricative	33%	44%	27%	72%
nasal/glide/liquid	18%	17%	20%	15%
h	1%	2%	4%	1%
Total N	64896	298	153	119

Table 4: Feature Class of Initial Phoneme in Fragments by Fragment Length

	Filled Pauses/Cue Phrases	Unfilled Pause
Fragment	16	264
Non-fragment	20	82

Table 5: Occurrences of Filled Pauses/Cue Phrases and Word Fragments

Pausal Juncture	Mean	Std Dev	N
Fluent Pause	513 msec	676 msec	1186
DI	334 msec	421 msec	346
Fragment	289 msec	377 msec	264
Non-fragment	481 msec	517 msec	82

Table 6: Duration of Silent DIs vs. Utterance-Internal Fluent Pauses

Repair Constituent	Tokens	%
Noun phrase	42	36%
Prepositional phrase	36	31%
Sentence	24	21%
Verb phrase	8	7%
Participial phrase	6	5%
Relative clause	1	0.9%

Table 7: Distribution of Syntactic Categories for Exact Constituent Repairs (N=117)

	No Boundary	*Boundary*	*Observed*
No Boundary	62.80%	3.73%	66.53%
Boundary	4.57%	28.90%	33.47%
Predicted	67.37%	32.63%	91.70%

Table 8: CART Predictions on Read TI Utterances, N=5471

tisri.10.snp

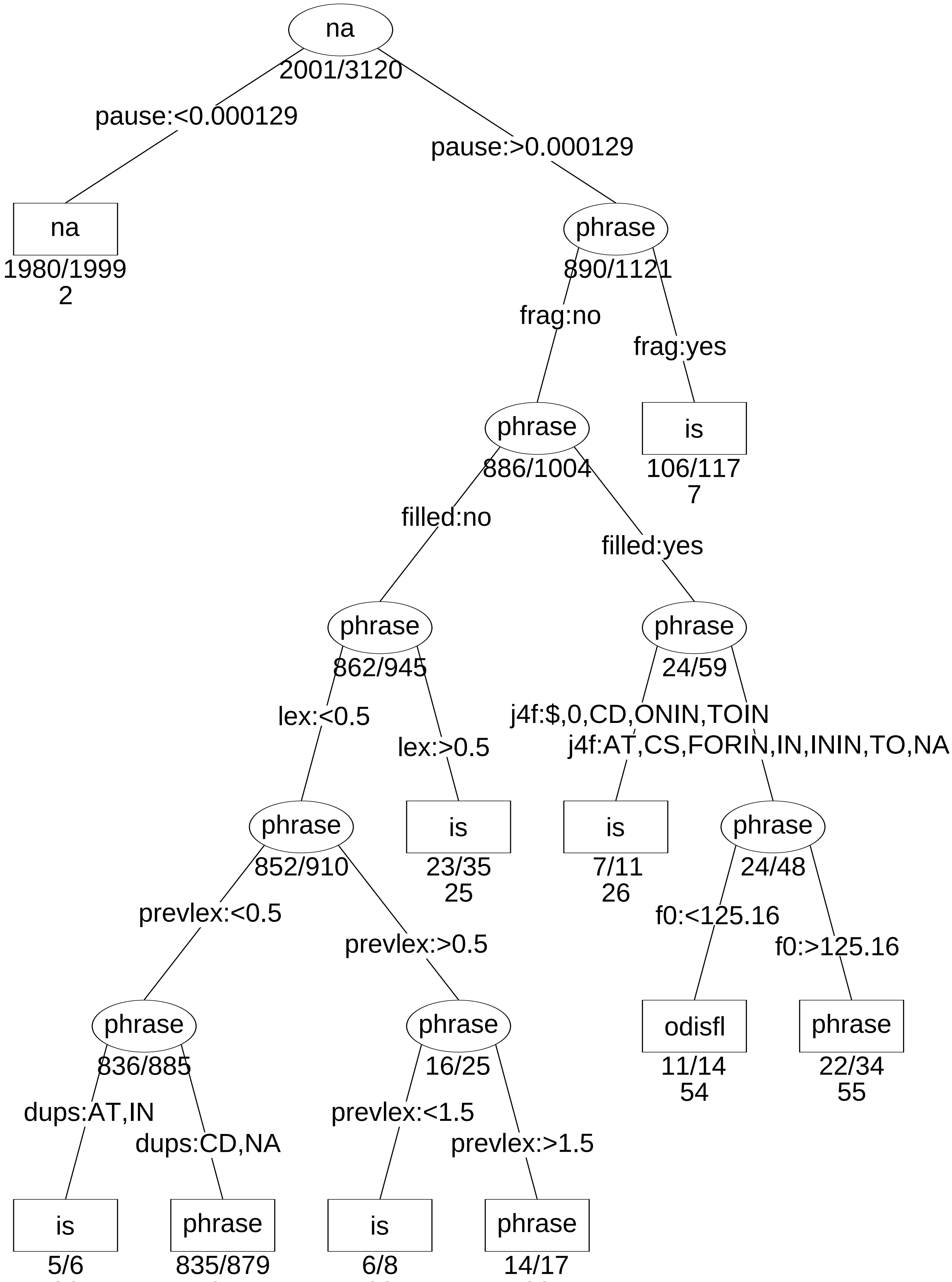
na
2001/3120
pause:<0.000129
pause:>0.000129
na
1980/1999
2
phrase
890/1121
frag:no
frag:yes
phrase
886/1004
is
106/117
7
filled:no
filled:yes
phrase
862/945
phrase
24/59
lex:<0.5
lex:>0.5
j4f:$,0,CD,ONIN,TOIN
j4f:AT,CS,FORIN,IN,ININ,TO,NA
phrase
852/910
is
23/35
25
is
7/11
26
phrase
24/48
prevlex:<0.5
prevlex:>0.5
f0:<125.16
f0:>125.16
phrase
836/885
phrase
16/25
odisfl
11/14
54
phrase
22/34
55
dups:AT,IN
dups:CD,NA
prevlex:<1.5
prevlex:>1.5
is
5/6
phrase
835/879
is
6/8
phrase
14/17

Figure Captions

FIG. V.. Predicting Disfluencies From Acoustic and Lexical Information