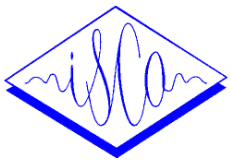

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

A SPEECH-FIRST MODEL FOR REPAIR IDENTIFICATION IN SPOKEN LANGUAGE SYSTEMS

Julia Hirschberg

AT&T Bell Laboratories
Murray Hill NJ 07974 USA

Christine Nakatani

Division of Applied Sciences, Harvard University
Cambridge MA 02138 USA

Abstract

Self-corrections or REPAIRS are often left unmodeled in current spoken language systems although they occur in about 10% of spontaneous utterances. We report on a study of acoustic-prosodic repair cues of potential use for repair identification, word fragment identification, and repair correction. The relative contributions of these and other text-based cues to repair identification were tested in a statistical model that achieved a precision rate of 91% and recall of 86%.

Introduction

Self-corrections or REPAIRS, which occur in about 10% of spontaneous utterances [6, 1], are often left unmodeled in spoken language systems. Yet repairs may cause recognition errors as in Example (1) or interpretation difficulties as in Example (2).[1]

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

Numerous text-based methods for handling repairs have been proposed that rely on the availability of accurate orthographic transcriptions to identify repairs, but little is known about how to support or complement such "text-first" approaches in the speech processing of repairs (cf. [15, 1]). Current recognition systems derive language models and lexicons primarily from fluent speech, treating disfluencies in training and recognition as noise.

To better understand the potential contributions of speech cues to repair processing, we studied the acoustic-prosodic features of 382 repairs from the ARPA Air Travel Information System (ATIS) database. We interpret our results within a "speech-first" framework for investigating repairs, the REPAIR INTERVAL MODEL (RIM), and test them in a statistical model that achieves a precision rate of 91% and recall of 86% for repair identification on a prosodically labeled corpus.

The Repair Interval Model

To support our investigation of acoustic-prosodic repair cues, 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 [9], 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*' or '*I mean*'), which indicate the speaker's recognition of his/her performance error. The REPAIR INTERVAL comprises the material intended to 'replace' the reparandum. It extends from the offset of the DI to the end of the correcting speech. In Example (3), 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.

(3) 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.

Labov [7] and Hindle [6] have hypothesized that an acoustic-phonetic edit signal, "a markedly abrupt cut-off of the speech signal" [6, p.123], occurs at the IS. Based on our analyses and on recent psycholinguistic experiments [10], this proposal appears too limited. Crucially, in RIM, we extend the notion of the edit signal to include any phenomenon which may contribute to the perception of self-interruption — including cues such as coarticulation phenomena, word fragments, interruption glottalization, pause, and prosodic cues which occur in the vicinity of the disfluency interval. As reconceived in the RIM model, the edit signal more generally marks the juncture between the reparandum and the repair intervals. This juncture has been shown to be useful for rule-governed correction strategies [6], further motivating the exploration of a broader range of acoustic-prosodic manifestations of the edit signal.

Acoustic-Prosodic Characteristics of Repairs

Our study of the acoustic and prosodic correlates of repair events as defined in the RIM framework extends a study reported in [12]. The current corpus consisted of 6,414 utterances from the ARPA Airline Travel and Information System (ATIS) database [11] collected at AT&T, BBN, CMU, SRI, and TI. 346 (5.4%) of these utterances produced by 122 speakers 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. Orthographic transcriptions of the utterances were prepared by ARPA contractors according to standardized conventions. The utterances were labeled at Bell Laboratories for

[1] Output in Example (1) is from the system described in [8]. 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.

word boundaries and intonational prominences and phrasing following Pierrehumbert's description of English intonation [13]. Also, each of the three RIM intervals and prosodic and acoustic events within those intervals were labeled. Speech analysis was done with Entropic Research Laboratory's WAVES software.

Identifying the Reparandum Interval

We did not identify any reliable cues for the onset of the reparandum, but we did find several cues at the reparandum offset. In our corpus, 73.3% (298/382) of all reparanda end in word fragments. 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. We identified the broad word class of the speaker's intended word for each fragment, where the intended word was recoverable. 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. Analysis of fragment length showed 91% of fragments were one syllable or less in length (40% single consonant, 51% single syllable).

Table 1 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 1 we see that single

Class	% of Words	% of Frags	% of One Syll Frags	% of One Cons Frags
stop	23%	23%	29%	12%
vowel	25%	13%	20%	0%
fric	33%	44%	27%	72%
nas/gl/liq	18%	17%	20%	15%
h	1%	2%	4%	1%
N	64896	298	153	119

Table 1: Feature Class of Initial Phoneme in Fragments by Fragment Length

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. Both the overall and repair distributions ($p<.001$,chistat = 32.88, df=4) and the single consonant and single syllable distributions ($p<.001$,chistat = 66.27, df=4) differ significantly.

Two additional acoustic-phonetic cues, glottalization and coarticulation, may aid fragment identification. Bear et al. [1] note that irregular glottal pulses sometimes occur at the reparandum offset. This form of INTERRUPTION 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, 29.8% of reparanda offsets are marked by interruption glottalization. Although interruption glottalization is usually associated with fragments, not all fragments are glottalized. In our database, 63.6% of fragments are *not* glottalized, and 10.5% of glottalized reparanda offsets are *not* 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.

We conclude that 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 may further improve the modeling of fragments.

Identifying the Disfluency Interval

In our corpus, pause fillers and cue phrases, which have been hypothesized as repair cues (cf. [2]), occur within the DI for only 9.4% (36/382) of repairs, and so cannot be relied on for repair detection. Interestingly, pause fillers and cue phrases occur significantly more often in non-fragment repairs than in fragment repairs ($p<.001$, chistat = 16.91, df=1).

We found another distinction between non-fragment and fragment repairs, namely the duration of pause following the IS. Table 2 shows the average duration of 'silent DI's (i.e. containing no pause fillers or cue words) compared to that of fluent (i.e. non-hesitation) utterance-internal silent pauses for the TI utterances. Overall, silent DIs are shorter than fluent pauses

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

Table 2: Duration of Silent DIs vs. Utterance-Internal Fluent Pauses

($p<.001$, tstat=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 nonfragments ($p<.001$, tstat=3.67, df=344). The fragment repair DI duration is also significantly shorter than fluent pause intervals ($p<.001$, tstat=5.20, df=1448), while there is no significant difference between non-fragment DIs and fluent utterances. So, DIs in general appear to be distinct from fluent pauses, and in particular the duration of DIs in might be exploited to identify cases of fragment repairs.

Identifying the Repair

Previous studies of disfluency have paid considerable attention to the vicinity of the DI but little to the repair offset. RIM analysis uncovered one general intonational cue that may be of use for repair *correction*, namely the prosodic phrasing of the repair interval.

First, we tested the hypothesis that repair interval offsets are marked by phrase boundaries, using the phrase prediction procedure reported by Wang & Hirschberg [16] to estimate whether

the phrasing at the repair offset was predictable according to a model of fluent phrasing.[2] We found that 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; 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.

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 [4, 5, 14], especially the relative inviolability of noun phrases [16], and the length of prosodic phrases [5]. On the one hand, we found phrase boundaries at repair offsets which occurred within larger NPs, as in Example (4), where it is precisely the noun modifier — not the entire noun phrase — which is corrected. (Prosodic boundaries in examples are indicated by '|'.)

(4) 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 Example (5), where the corresponding fluent version of the repair interval is predicted to contain four phrases.

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

In both cases above, 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.[3] We found three major classes of phrasing behaviors. First, as noted above, for 43% (165/382) of repairs, the repair offset we had initially identified 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 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. We propose that this class of repairs be identified as constituent repairs, rather than the lexical repairs we had initially hypothesized. For the majority of these constituent repairs (77%, 117/151), the repair interval contains a well-formed syntactic constituent (see Table 3). If the repair interval does *not* form a syntactic constituent, it is most often an NP-internal repair (74%, 25/34). 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 (4)), phonetic errors, word insertions, and syntactic reformulations.

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

Table 3: Distribution of Syntactic Categories for Exact Constituent Repairs (N=117)

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, which are prevalent in the ATIS corpus due to its 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 prosodic cue may aid repair processing by delimiting the interval over which correction strategies may operate.

Predicting Repairs from Acoustic and Prosodic Cues

We next investigated the predictive power of our characterization of repairs derived from RIM analysis on the ATIS corpus. 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 [12] as training data; the additional 202 repair utterances (containing 223 repair instances) were used for testing. We attempted to distinguish repair IS from fluent phrase boundaries (collapsing major and minor boundaries), non-repair disfluencies (marked independently of our study and including all events with some phonetic indicator of disfluency which was not involved in a self-repair) and simple word boundaries. Our data points included every position between two words and are represented below as ordered pairs $<w_i,w_j>$, where w_i and w_j represent the lexical items to the left and right of the potential IS respectively.

For each $<w_i,w_j>$, we examined the following features as potential IS predictors: duration of pause between w_i and w_j; occurrence of word fragment(s) within the $<w_i,w_j>$ interval; occurrence of a filled pause in the $<w_i,w_j>$ interval; energy peak within w_i (absolute as well as normalized for utterance); amplitude of w_i relative to w_{i-1} and to w_j; absolute and normalized f0 of w_i; f0 of w_i relative to w_{i-1} and to w_j; and w_i's accent status (accented or deaccented). We also simulated simple pattern matching strategies, to see how acoustic-prosodic

[2] 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 (N=63).

[3] Hindle [6] defines a typology of repairs and associated correction strategies: 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; or (3) surface level or lexical repairs, in which identical strings appear adjacent to each other.

cues might interact with lexical cues in repair identification, looking at distance in words of w_i from beginning and end of the utterance; total number of words in utterance; whether w_i or w_{i-1} recurred in the utterance within a window of three words after w_i; a part-of-speech window of four around the potential IS; and whether, if function words, w_i and w_j shared the same part-of-speech (e.g. PREP PREP).

We trained prediction trees using Classification and Regression Tree (CART) techniques [3] given these features. The resulting tree was then used to predict boundary locations in our test set. This procedure identified 192 of the 223 repairs observed in the test set, with 19 false positives, representing a recall of 86.1% and precision of 91.2%. All repairs were 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 p.o.s. information; b) lexical matching across the DI; and c) duplication of p.o.s. across the DI. Thus, the usefulness of combining general acoustic-prosodic constraints with lexical pattern matching techniques as a strategy for repair identification appears to gain some support from these data.

Discussion

The Repair Interval Model has guided our study of the acoustic-prosodic features of repairs in spontaneous speech. Our results indicate that self-interruption, may be conveyed by a number of different cues, including word fragmentation, glottalization, coarticulatory effects preceding silent pauses, and the duration of the disfluency interval itself. We identified several features to aid in fragment identification, such as the distributions of fragments by length and by initial phoneme. We also determined that repair intervals may differ from fluent speech in their characteristic prosodic phrasing, and noted the role prosody might play in delimiting the repair interval for correction strategies. Although the full integration of these acoustic-prosodic findings with existing proposals for repair detection and correction remains to be done, a first step in this direction was taken by our predictive modeling of repairs in the ATIS domain using CART analysis. Larger corpora must be examined, but our results of 86% recall and 91% precision, while preliminary, show that sufficient cues may exist in the vicinity of the DI to identify the majority of repairs in a local manner. Our study should contribute to the development of repair processing models that directly exploit the speech signal as a source of repair cues and that productively integrate such cues with other established text-based cues to repairs.

References

[1] J. Bear, J. Dowding, and E. Shriberg. 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, 1992. Association for Computational Linguistics.

[2] E. R. Blackmer and J. L. Mitton. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39:173–194, 1991.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove CA, 1984.

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

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

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

[7] W. Labov. On the grammaticality of everyday speech. Paper Presented at the Linguistic Society of America Annual Meeting, 1966.

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

[9] W. Levelt. Monitoring and self-repair in speech. *Cognition*, 14:41–104, 1983.

[10] R. J. Lickley, R. C. Shillcock, and E. G. Bard. 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 1991. Eurospeech-91.

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

[12] C. Nakatani and J. Hirschberg. A speech-first model for repair detection and correction. In *Proceedings of the 31st Annual Meeting*, Columbus, OH, 1993. Association for Computational Linguistics.

[13] J. B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, September 1980. Distributed by the Indiana University Linguistics Club.

[14] E. O. Selkirk. 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, 1984. TAPIR.

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

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