Proceedings of DiSS'03, Disfluency in Spontaneous Speech Workshop, 5–8 September 2003, Göteborg University, Sweden.
Robert Eklund (ed.), Gothenburg Papers in Theoretical Linguistics 90, ISSN 0349–1021, pp. 91–94.

Prosodic features of four types of disfluencies

Guergana Savova† & Joan Bachenko‡

† Medical Informatics Research, Mayo Clinic, Rochester, Minnesota, USA
‡ Linguistech Consortium, Oxford, New Jersey, USA

Abstract

We present a corpus-based approach for using intonation and duration to detect disfluency sites. The questions we aim to answer are: What are the prosodic cues for each disfluency type? Can predictive models be built to describe the relationship between disfluency types and prosodic cues? Are there correlations between the reparandum onset and offset and the repair onset and offset? Is there a general prosodic strategy? Our findings support four main hypotheses: 1) The Combination Rule: A single prosodic feature does not uniquely identify disfluencies or their types. Rather, it is a combination of several features that signals each type. 2) The Compensatory Rule: If there is an overlap of one prosodic feature, then another cue neutralizes the overlap. 3) The Discourse Type Rule: Prosodic cues for disfluencies vary according to discourse type. 4) The Expanded Reset Rule: Repair onsets are dependent on reparandum onsets and reparandum offsets. The limitation of the current study is the relatively small corpus size. Further testing of our proposed hypotheses is needed.

1. Introduction and background

Disfluencies have been studied from different angles – across several languages, domains (human–human and human–computer interaction), degrees of planning (spontaneous vs. read speech), from production or perception point of view, from a theoretical or application-oriented perspective. Our goal is to add to this body of research by reporting results on a corpus of semi-spontaneous, medical dictation speech by investigating basic prosodic characteristics (duration and intonation) of four disfluency types.

Prosodic studies of disfluency center on three acoustic features: intonation (fundamental frequency or F0), segment duration and pause duration. Lickley [6] shows in a controlled study that humans recognize a disfluency by the end of the first correct word even before accessing the semantic and syntactic information. Lickley comes to the conclusion that humans use prosodic information to detect disfluent speech. Oviatt et al. [9] propose a comprehensive prosodic model for disfluencies in human-computer interaction, the Computer-Elicited Hyperarticulated Model (CHAM). CHAM predicts that when the overall error rate of the system is low, the correction of the misrecognized word will involve only durational changes. When the overall error rate of the system is high, the prosodic characteristics of the correct word will have durational, articulatory (hyperarticulation), intonational and amplitude changes from its first occurrence. Hindle [2] relies on an abrupt cut-off signal to detect disfluencies and trigger his parser for disfluency correction.

The terminology used in our study follows the Repair Interval Model proposed by Nakatani & Hirschberg [7]. Each repair interval consists of three parts: a reparandum, which is the part to be repaired; a repair site, which provides the "new" material that corrects the reparandum; and the disfluency site, which contains any silences and filled pauses that may occur between the reparandum and repair site.

We propose four main hypotheses for the investigation of prosodic characteristics: 1) The Combination Rule: A single prosodic feature does not uniquely identify disfluencies or their types. Rather, it is a combination of several features that signals each type. 2) The Compensatory Rule: If there is an overlap of one prosodic feature, then another cue neutralizes the overlap. In other words, if one prosodic feature is strongly indicated suggesting multiple possibilities for prosodic boundaries (e.g. prolongation occurs at utterance boundaries, but also at utterance internal repetition sites), then another feature will disambiguate the final choice (e.g. prolongation with an utterance final tone indicates an utterance boundary vs. prolongation with sustained or repeated contour indicates an utterance internal repetition). 3) The Discourse Type Rule: Prosodic cues for disfluencies vary according to discourse type, e.g. human–human vs. human–machine interaction. 4) The Expanded Reset Rule (based on Pike [12]): Repair onset F0 values are dependent on the F0 values of reparandum onsets and reparandum offsets.

The main research questions suggested by the hypotheses are: What are the prosodic cues for each disfluency type? Can predictive models be built to describe the relationship between disfluency types and prosodic cues? Are there correlations between the reparandum onset and offset and the repair onset and offset? Is there a general prosodic strategy or is it discourse-dependent?

2. Method

2.1. Corpus description and disfluency tagging

Our study is data-driven and based on a corpus collected by Linguistic Technologies, Inc. (LTI), a company that applied automatic speech recognition to medical dictations. There are 21 talkers yielding 32,122 words approximately evenly distributed among the talkers. The speaking style is classed as quasi-spontaneous. The physicians have notes and templates to follow but fill in template sections with spontaneous discourse.

To categorize disfluencies, we use the classification scheme described in Page [10] motivated by the two criteria: the categories must be mutually exclusive and must allow for cross-comparison and further subclassification. Three undergraduate linguistics students from University of Minnesota tagged the disfluencies. Sites where the classification decisions differed were discussed and a final tag was agreed upon. The disfluency types are:

- Exact repetitions (type 1): single or multiple word repetitions separated optionally by filled pauses, silence, editing expressions, or any combination of these, e.g. "the um | the" (88 sites).
- Exact substitution (type 2): single or multiple word substitutions, separated optionally by silence, filled pauses, editing expressions, or any combination of these, e.g. "five correction | seven" (182 sites).
- Repetition and substitution (type 3): substitution with repeated material to the left or the right, e.g. "does not | did not" (72 sites).

- Repetition and insertion (type 4): repetitions with a new word inserted before or medially, e.g. “to clean | to try to clean.” (20 sites).
- Repetition and deletion (type 5): repetitions with a word omitted either at the start of the repeat or medially, e.g. “no spotting dysuria or abnormal | no spotting or dysuria” (4 sites).

Unlike other research [1], fragments are not classified as a completely separate group; instead, they are treated as words. We report results on the first 4 disfluency types as type 5 occurs infrequently in our corpus. Also, our study excludes sites with editing expressions, e.g. “five correction | seven”.

Exact substitutions were further broken down into subgroups by 3 subclassification features to allow comparisons with Levelt & Cutler [5], a study suggesting that syntactic and phonetic errors do not receive any prosodic marking, but semantic errors form a separate group and tend to be prosodically marked:

- Feature 1 – What does the repair fix?
  - Pronunciation, e.g. “sci- | scaling” (109 sites)
  - Semantics, e.g. “throat | lungs” (32 sites)
  - Syntax, e.g. “he | his” (11 sites)
  - Semantics/syntax, e.g. “is appa- | somehow got lost” (13 sites)
- Feature 2 – Is there a fragment at the reparandum?
  - Yes, e.g. “sci- | scaling” (108 sites)
  - No, e.g. “a | what” (57 sites)
- Feature 3 – how can the reparandum be described in regard to the repair?
  - Mispronunciation, e.g. “ma- | mycitracin” (31 sites)
  - Repeat, e.g. “ec-“ | “exercises” (42 sites)
  - Semantic error, e.g. “throat| lungs” (51 sites)
  - Syntactic error, e.g. “one | once” (15 sites)
  - Semantic/syntactic error, e.g. “talk | thinking” (13 sites)
  - Needed elaboration, e.g. “ec- | low back exercises” (12 sites)
  - Multiple corrections needed, e.g. “she’s had ah ah he sen- | ah she is” (1 site)

2.2. Research variables

A number of studies investigate duration and F0 contours as the most salient prosodic features for disfluency modeling. Our study focuses on these features as well and describes them in the context of disfluency types 1–4 — exact repetitions, exact substitutions, repetitions with substitutions and repetitions with insertions.

The raw duration values are normalized by two formulas and comparisons are done with normalized values:

$$norm\_value1 = \frac{raw\_duration - mean}{st.deviation}$$

$$norm\_value2 = \frac{raw\_duration}{mean}$$

Small size samples (N<15) were excluded from the study as the standard deviation for those would exhibit a large spread. Segmental durations were obtained by force-aligning audio files with their respective text using the speech recognizer developed at Entropic Cambridge Research Laboratory. The alignment was hand-checked for correctness.

F0 tracks and values were extracted using the Entropic XWAVES+ software. All the values for the entire repair interval were checked for spurious doubling and halving, and, where needed, the values were hand-corrected. Samples with vocal fries were excluded from the study because their F0 values are unusually low [8]. Glottalization, on the other hand, does not appear to be associated with a sustained decrease in F0 [3]. Sites with word-medial voiceless fricatives were excluded from the analyses as they introduce spurious F0 values.

The F0 contours (onset, max or min, offset) are taken over the reparandum and repair words only, not the entire sentence or prosodic phrase. The contours are represented as sequences of low (L) and high (H) tones, based on the F0 values for the given word. The contours reflect the onset, peak, valley and offset F0 from the word in focus. The contours are a simplified version of Pierrehumbert’s [11] intonational system and are taken over the word excluding the accented (*) tone and phonological analyses of the tones. Thus, only the overall F0 movement over the repair interval is recorded. The intonation data is presented in graphs regardless of durational characteristics at equal intervals between points. For the intonation analyses, disfluencies were further subdivided according to the presence/absence of silences at the left and/or right edge of the repair interval and the disfluency site. Because of the shrinking sample size, intonation results are reported only for exact repetitions and exact substitutions.

3. Results and discussion

The results for the four disfluency types are graphed in Figures 1, 2, 3 and 4. Due to space limitations we do not report the results for the disfluency site silences and their embedding in the discourse hierarchy as compared to silences at utterance boundaries and boundaries at bigger discourse segments. These can be found in Savova [13]. The current figures represent the tendencies for exact repetitions and exact substitutions in two layers. The lower part reports segmental durations for reparanda, disfluency sites and repair sites. Each group above the fluent speech band shows scores significantly different from the fluent speech means ($p<0.05$). Each group represented by a separate curve exhibits results significantly different from the other groups ($p<0.05$). The upper part of the diagram reports the generalized tendency of the intonational contours along with the F0 range in Hz and the starting point of the repair as compared to the reparandum onset. Due to the small sample size for the intonational analyses, we report only durational results for type 3 and type 4 disfluencies. Detailed scores from the statistical tests can be found in Savova [13]. Correlations between reparandum F0 onsets and offsets and repair F0 onsets and offsets are strong (range is 0.62–0.99) providing a basis for predictive statistical models to be included in a disfluency detection algorithm. The lower correlation values are for repair intervals surrounded by silences. The correlation results need to be compared against fluent speech relations to find the uniqueness when a disfluency is present. Such a comparison is left for future research.

Our data supports only partially Pike [12] as there are tokens whose repair starts at F0 values lower than those of the reparandum onsets. The strong correlations between the reparandum offset values and repair onset/offset values hint of dependencies that go beyond reparandum onset values, which supports our proposed *Expanded Reset Rule*.

Exact repetitions and exact substitutions show unique intonation patterns – matching intonational contour for the repetitions and a final low tone for the substitutions. According to the central tendencies presented in the figures, the substitution repair has larger F0 range in comparison to its reparandum and the repetition (type 1) repair.

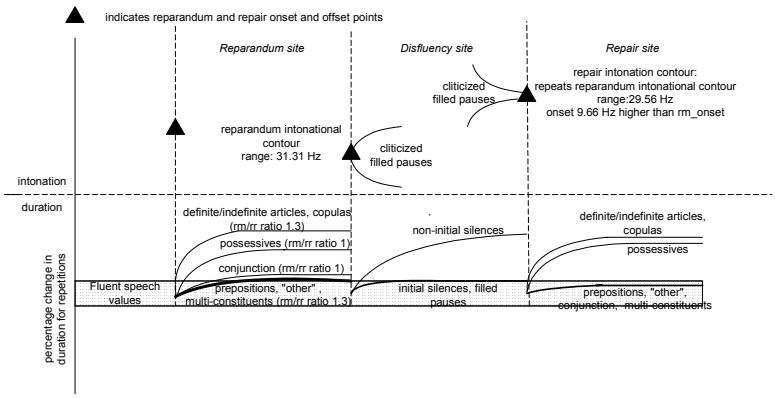

Figure 1: Summary of the results for exact repetitions (type 1) – duration and intonation.

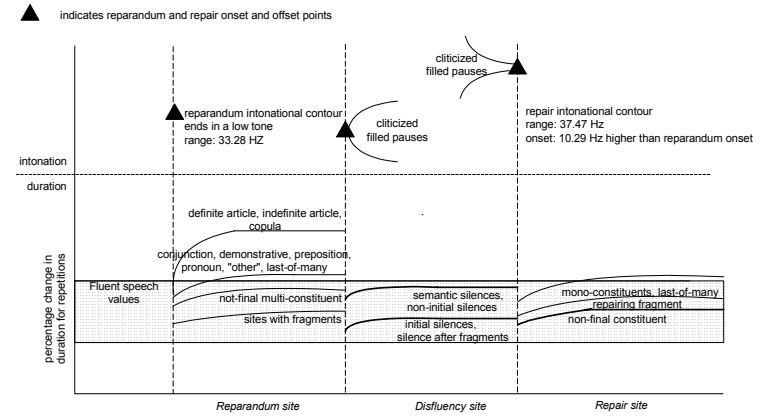

Figure 2: Summary of the results for exact substitutions (type 2) – duration and intonation.

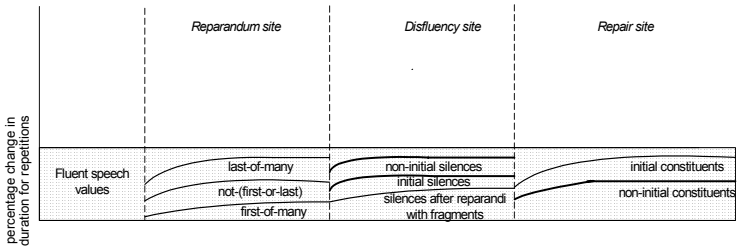

Figure 3: Summary of the results for repetitions with substitutions (type 3) – duration.

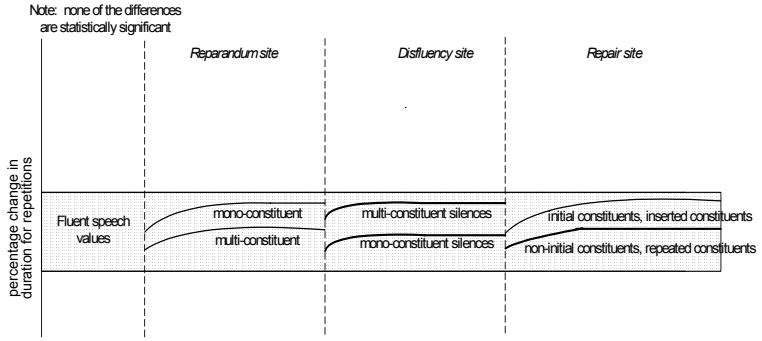

Figure 4: Summary of the results for repetitions with insertions (type 4) – duration.

Larger F0 ranges are one of the methods for prosodic marking (Pierrehumbert [11]). Durational analyses show that mono-constituent reparanda are consistently longer than reparanda consisting of many words (multi-constituent reparanda) across the disfluency types studied. Fragments have a contracting effect on the duration of the reparanda and of the disfluency site components.

Comparing by subclassification features for exact substitutions did not provide consistent results, thus only partial support for Levelt & Cutler's proposal [5]. Our result is in accord with Hokkanen [4] who also reports only partial support for Levelt & Cutler [5]. It must be noted that the methods in the studies differ – [5] uses subjective judgements for prosodic marking, while [4] and the current study use F0 measures. Future research must include converging methodologies.

We found initial support for the *Discourse Type Rule*. Oviatt's CHAM [9] is only partially supported by our study giving additional ground to believe that disfluencies differ by type of discourse – Oviatt's study was done on forced errors in a human-computer interaction, while our study deals with spontaneous disfluencies in monologue-like speech.

There is good initial support for the *Compensatory Rule* – no two features are overly emphasized as that might lead to confusion with other discourse segment boundaries. For example, in the case of definite/indefinite article repetition, there is considerable prolongation, but the F0 fluctuations are moderate which rules out a possible utterance boundary.

The *Combination Rule* also found good initial support. One signal or "an abrupt cut-off" as Hindle [2] defined it, does not uniquely identify the right edge of the reparandum. Hence, our proposal is to exploit prosodic combinations as they naturally occur in speech. For example, mono-constituent repetition reparanda are prolonged, but so are mono-constituent substitutions. Repetition reparanda have a falling contour, but so do substitution reparanda. The feature that is closest to being type unique is the duration of the disfluency site silence, but it is not present at all sites. If layers of prosodic information and placement within the overall prosodic discourse structure are combined, the uniqueness of disfluency prosodic characteristics might emerge.

4. Conclusions

Future research will investigate the inclusion of the disfluencies in the overall discourse hierarchy. Based on the combined results, our goal is to offer a computationally viable algorithm for disfluency detection via prosodic characteristics. For that, we need to further study the relations between fluent and disfluent speech and expand the corpus size which is the limitation of the current study.

5. References

[1] Clark, Herbert & Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology.* vol. 37, pp. 201–242.

[2] Hindle, Donald. 1983. Deterministic parsing of syntactic non-fluencies. *Proc 21st ACL*. pp. 123–128.

[3] Hirschberg, Julia. 2000. A corpus-based approach to the study of speaking style. *Prosody, Theory and Experiment.* Kluwer Academic Publ. ISBN 0-7923-6579-8.

[4] Hokkanen, Tapio. 2001. Prosodic marking of self-repairs. *Proc DiSS'01*. University of Edinburgh. Sept 2001. pp. 37–40.

[5] Levelt, Willem & Anne Cutler. 1983. Prosodic marking in speech repairs. *Journal of Semantics*, vol. 2, no. 2, pp. 205–217.

[6] Lickley, Robin. 1994. Detecting disfluency in spontaneous speech. PhD dissertation, University of Edinburgh.

[7] Nakatani, Christine H. & Julia Hirschberg. 1994. A corpus-based study of repair cues in spontaneous speech. *JASA*, vol. 95(3), pp. 1603–1616.

[8] Olive, Joseph, Alice Greenwood & John Coleman. 1993. *The acoustics of American English speech: a dynamic approach.* New York: Springer.

[9] Oviatt, Sharon, Margaret MacEachern & Gina-Anne Levow. 1998. Predicting hyperarticulate speech during human–computer error resolution. *Speech Communication*, vol. 24, pp. 1–23.

[10] Page, Sherri. 1999. Use of a postprocessor to identify and correct speaker disfluencies in automated speech recognition for medical dictations. *Proc. DiSS'99.* San Francisco. July 1999. pp. 27–30.

[11] Pierrehumbert, Janet. 1980. The phonology and phonetics of English intonation. PhD dissertation. MIT.

[12] Pike, Kenneth. 1945. *The intonation of American English.* Ann Arbor, MI: University of Michigan Press.

[13] Savova, Guergana. 2002. *Disfluencies, prosody and discourse in quasi-spontaneous speech*. Unpublished PhD dissertation, University of Minnesota.