Prosodic parallelism as a cue to repetition disfluency

Jennifer Cole, Mark Hasegawa-Johnson, Chilin Shih, Heejin Kim, Eun-Kyung Lee, Hsin-yi Lu, Yoonsook Mo, Tae-Jin Yoon

University of Illinois at Urbana-Champaign, USA

Abstract

Repetition disfluencies are among the most frequent type of disfluency in conversational speech, accounting for over 20% of disfluencies, yet they do not generally lead to comprehension errors for human listeners. We propose that parallel prosodic features in the REP and ALT intervals of the repetition disfluency provide strong perceptual cues that signal the repetition to the listener. We report results from a transcription analysis of repetition disfluencies that classifies disfluent regions on the basis of prosodic factors, and preliminary evidence from F0 analysis to support our finding of prosodic parallelism.

1. Acoustic-prosodic correlates of disfluency

Disfluency occurs in spontaneous speech at a rate of about one every 10-20 words, or 6% per word count [17], yet this interruption of fluent speech does not generally lead to comprehension errors for human listeners. Recent research has shown that important cues to disfluency can be found in the syntactic and semantic structures conveyed by the word sequence, and in the phonological and phonetic structures signaled by acoustic features local to the disfluency interval. These cues identify the components of the disfluent region--- the reparandum (REP), edit phrase (EDIT), and alteration (ALT)--- and their junctures. Work on automatic disfluency detection has shown that the most successful approach combines both lexical and acoustic features, with explicit models of the lexical-syntactic and prosodic features that pattern systematically with disfluent intervals [1,6].

Of the acoustic-prosodic correlates of disfluency, the post-reparandum pause (filled or unfilled) has been studied the most extensively. Nakatani & Hirschberg's [12] detailed acoustic and classification studies examine duration, F0 and energy, and also report unusual patterns of lengthening, coarticulation, and glottalization near the interruption point of a disfluency. In this paper we examine the nature of prosodic correlates of disfluency in the characteristic patterns of F0, duration and energy that identify and distinguish among various types of disfluency involving word repetition.

There are distinct types of disfluency that can be characterized in terms of their form and function. Shriberg [16, 17] classifies the disfluencies of the Switchboard corpus into six categories: filled pause ("uh" and "um"), repetition (of one or more words, without correction), substitution (repetition of zero or more words, followed by the correction of the last word in the disfluent interval), insertion, deletion, and speech error. Other work identifies abandonment (fresh start) disfluencies, in addition [6,11,18]. These distinct types of disfluency may be caused by different psychological processes. Levelt [9] suggests that corrections of a single word may result from monitoring of the phonetic plan, while corrections that involve repair or abandonment of an entire phrase may result from monitoring of the pre-syntactic message. Clark and Fox Tree [3] and Clark & Wasow [4] propose a different psychological account for filled pause and repetition disfluencies. In these accounts filled pauses like "uh" and "um" are phonological words that are used by the speaker to signal a delay in the preparation of the upcoming speech. Repetition disfluencies occur when the speaker makes a premature commitment to the production of a constituent, perhaps as a strategy for holding the floor, and then hesitates while the appropriate phonetic plan is formed. The continuation of speech is marked by "backing up" and repeating one or more words that precede the hesitation, as a way of restoring fluent delivery. Henry and Pallaud [7] support the findings of Clark & Wasow [4] by demonstrating that morphological, syntactic, and structural features strongly differentiate repetition disfluencies from word fragment disfluencies. Clark & Wasow [4] note that repetition disfluencies are four times as common as repair disfluencies; they suggest that a small number of repetition disfluencies may be "covert repairs" [9], but that most repetitions are more closely related to filled pause disfluencies than to speech repairs.

The acoustic-prosodic features that serve to cue disfluency vary according to the type of disfluency. Levelt & Cutler [10] observe a contrastive emphasis on the repair segment of an error-correcting disfluency, manifest in increased F0, duration and amplitude. Shriberg [15] and Plauché & Shriberg [13] find that F0 contours, word durations, and the distribution of pauses serve to differentiate among three types of repetition disfluencies. Shriberg [15] describes repetition disfluencies that signal covert repair as having a characteristic reset of the F0 contour to a high, phrase-initial value at onset of the alteration. Similarly, Savova and Bachenko [14] propose an "expanded reset rule," according to which "alteration onsets are dependent on both reparandum onsets and reparandum offsets," echoing the observation of Shriberg [15] that when speakers modify the duration of a repeated word in a repetition disfluency, "they tend to do so in a way that preserves intonation patterns and local pitch range relationships."

In our study of prosody and disfluency in the Switchboard corpus of conversational telephone speech, we observe parallelism in the prosodic features of the REP and ALT phases as characteristic of some repetition disfluencies. Highly similar F0 patterns express a parallel intonation structure that cues the relationship between the REP and ALT for the majority of repetition disfluencies we have observed. We propose an extended typology of repetition disfluencies in this paper, based on prosodic comparison of REP and ALT. Section 2 describes the methods of our transcription study of disfluency in Switchboard, and section 3 presents frequency data on five types of repetition disfluency that are prosodically distinguished based on a comparison of the prosodic features of the REP and ALT intervals. Section 4 reports on preliminary quantitative evidence from F0 data to support our analysis based on perceptual transcription.

2. Method

2.1. Corpus

Switchboard is a corpus which consists of 2500 spontaneous informal telephone conversations [5]. We selected 70 sound files from those conversations, representing 58 different speakers. Within each file we used a random process to excerpt a two minute sound segment. These short files were transcribed for disfluency intervals by the authors, all of whom are trained in acoustic phonetics with prior experience in prosodic transcription using ToBI annotation conventions. 3 transcribers labeled disfluencies for the entire two-minute duration of 10 files each (for a total of 60 minutes of speech) and 5 transcribers labeled only for the first talker turn of duration between 3-60 ms. in each of 10 files (for approximately 25 minutes of speech). All eight labelers participated in a series of three group training sessions to assure consistency of labeling criteria, and two group sessions were held for the resolution of problem cases raised by individual labelers.

2.2. Labeling Criteria

Disfluencies are classified by their function into two types, hesitation and repair. These functional categories divide into several subtypes based on lexical and prosodic form. Hesitation disfluencies are classified as repetition, lengthening, silent pause and filled pause. Repair disfluencies are classified as error correction and abandonment. Classification was based on lexical, syntactic, and prosodic factors. Lexical factors are the presence of a repeated word, an error-correcting word substitution, or a filled-pause phrase like “um” or “ah”. Syntactic criteria were used to identify instances of phrase abandonment followed by fresh restart and to identify the REP-ALT correspondence in error-correction. Prosodic factors were used to identify lengthening, and provided additional evidence for some cases of error correction (with prosodic emphasis on ALT) and abandonment (with truncation of an intonational tune at the abandoned edge). Labeling was done on the basis of listening and visual inspection of the waveform, spectrogram, F0 and intensity contours, using Praat [2]. The disfluency labels were entered on two tiers in the TextGrid associated with each wave file, and disfluency intervals (REP, EDT, ALT) were aligned with the beginnings and endings of the associated word intervals. Table 1 shows the typology of disfluencies by function and form and the labeling conventions used.

Table 1. Typology of Disfluencies and Labeling Convention

Type of Disfluency		Labeling			
		1st Tier	2nd Tier		
Hesitation	Repetition	hesi-r	REP	EDT	ALT
			REP	ALT	
	Lengthening	hesi-l			
	Silent Pause	hesi-s			
	Filled Pause	hesi-f			
Repair	Error Correction	repair-e	REP	EDT	ALT
			REP	ALT	
	Abandonment	repair-a	REP	EDT	
			REP		

Labels on the first disfluency tier identify the type of disfluency (e.g., hesi-r for Hesitation Repetition), while the components of complex disfluencies were individually segmented on the second disfluency tier. A complex disfluency always includes a reparandum (REP), and may also include an edit phase (EDT) and an alteration (ALT). Hesitation Repetition disfluency labeling is illustrated in Figure 1. The hesi-l label marks hesitation lengthening that can not be attributed to prosodic phrase-final lengthening given based on tonal evidence and perceived disjuncture. In addition, hesi-s denotes a sentence internal silence that interrupts an otherwise fluent phrase, and hesi-f marks an independent occurrence of filled pause expressions such as “um”, “uh”. For the repair category, repair-e marks an error followed by a self-correction (e.g. “he can stri- he can swing”) and repair-a denotes a semantic and syntactic abandonment of the phrase (e.g. “they you know you can’t live in Dallas”).

…the kids	instead of	[sil]	instead of	teaching…
	hesi-r			
	REP	EDT	ALT	

Figure 1: TextGrid tiers for a Hesitation Repetition disfluency [Switchboard file: SW03719A.wav]

The EDT label marks the occurrence of filled pauses, silent pauses and editing expressions (e.g.,“I mean, you know”) between REP and ALT.

The Hesitation Repetition disfluencies were further broken down into five sub-classes based on comparison of prosodic features between REP and ALT. These five sub-classes, listed in Table 2, were proposed on the basis of our earlier exploratory analyses with Switchboard samples; the present study was designed to test the adequacy and acoustic correlates of the proposed classification scheme. Data used in the exploratory analysis were not included in the present study. Prosodic features were assessed on the basis of listening in conjunction with visual inspection of the F0 and intensity contours, spectrogram and waveform. Repetitions in which the ALT and REP were judged to have highly similar prosodic patterns, with identical intonation features in a ToBI transcription, were assigned the label suffix ‘-same’. An example pitch track from a Repetition-Same disfluency is shown in Figure 2. The ‘-fp’ label was used for cases where the ALT interval had prosody characteristic of a filled pause: low intensity, and low, flat F0, with reduced consonant or vowel articulations. The ‘-ip’ label was used to label cases where the REP was perceived as the final word in a well-formed intermediate phrase, based on the F0 contour and perceived disjuncture between REP and the onset of ALT. The ‘-exaggerated’ label was applied to examples in which the ALT displayed a similar but exaggerated version of the prosodic pattern of the REP, typically with increased duration, intensity and higher F0 values. In many cases these examples would receive the same ToBI transcription for REP and ALT, with differences in F0 scaling. Finally, the label ‘-change’ was used for examples where the ALT differed prosodically from the REP in its accentuation (different type or location of accent, or presence vs. absence of accent). Change Repetitions sounded like error corrections, where the correction was at the level of pragmatic meaning expressed through accent, rather than at the level of word or syntactic meaning. For all disfluency types, the REP interval was further identified as ending in a word fragment (frag), or a complete word (nonfrag). These labels were abbreviated as indicated in Table 2, e.g., hesi-r-1a indicates a Hesitation Repetition with the same pitch pattern on REP and ALT and with no truncation of the final word in the REP phase.

Table 2. **Types of Repetition: Prosodic Classification**

Hesitation-Repetition		label
1. hesi-r-same	a. nonfrag	hesi-r-1a
	b. frag	hesi-r-1b
2. hesi-r-fp	a. nonfrag	hesi-r-2a
	b. frag	hesi-r-2b
3. hesi-r-ip	a. nonfrag	hesi-r-3a
	b. frag	hesi-r-3b
4. hesi-r-exaggerated	a. nonfrag	hesi-r-4a
	b. frag	hesi-r-4b
5. hesi-r-change	a. nonfrag	hesi-r-5a
	b. frag	hesi-r-5b

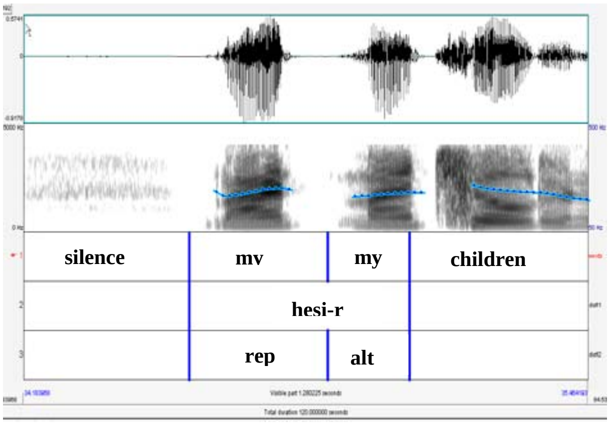

Figure 2. Example of highly similar F0 tracks on REP and ALT (my…my) in Repetition-Same disfluency: "[sil] my my children…" [Switchboard file: SB03633b]

3. Results

Table 3 provides the number of tokens of each type of disfluency labeled in the corpus, pooling data from all labelers. The most frequent type of disfluency in this corpus is Hesitation, with Silence the most frequent sub-type. Repetitions and Filled Pauses are also frequently occurring Hesitation types. Among Repair disfluencies, Abandonment is fairly common, while error correction and lengthening are infrequent.

Table 3. Distribution of the types of disfluency

		Frequency	Percentage
Hesitation	Silence	267	32.05 %
	Repetition	183	21.96 %
	Filled pause	182	21.84 %
	Lengthening	46	5.52 %
Repair	Abandonment	103	12.36 %
	Error correction	52	6.25 %
Total		833	

Table 4 presents the total number of REP, EDT and ALT intervals, automatically extracted from Hesitation Repetition and Repair Error Correction disfluencies in the transcription files. The Repair Error Correction disfluencies are not included in our perceptual comparison of REP and ALT for prosodic parallelism, but are included in the F0 data presented in section 4. The number of REP and ALT intervals are not equal, due to the occurrence of multiple repetition tokens that contain more than two instances of the repeated word (e.g., "I I uh I tried to…"). For multiple repetitions all but the non-final repetition are coded as independent REP intervals, with the final repetition coded as ALT.

Table 4. Distribution of REP, EDT, and ALT for Hesitation Repetition And Repair Error Correction

	REP	EDT	ALT
Repetition	201	81	182
Error Correcction	53	19	52
Total	254	100	234

The distribution of disfluencies in our corpus over the 10 sub-classes is shown in Table 5. We observe that word fragments are not common at the end of the REP phase for repetition disfluencies. Repetitions in which REP and ALT have the same prosody (hesi-r-same) are the most numerous, and are more than twice as frequent as repetitions that mimic filled pauses, cross intermediate phrase boundaries, or display exaggerated prosody on ALT, all of which occur with roughly equal frequency. Repetitions that exhibit a prosodic change on the ALT are the least frequent, which missors the low frequency of error corrections in Repair disfluencies (Table 4).

Table 5. Number of Hesitation Repetition examples by sub-class.

	hesi-r-same	hesi-r-fp	hesi-r-ip	hesi-r-exag	hesi-r-change
nonfrag	56	25	21	23	11
frag	6	2	1	5	3

4. F0 Analysis

F0 values were compared between REP and ALT as an empirical measure of intonational similarity. This section describes the method for extracting smoothed F0 contours, time normalization, and a measure of F0 contour difference.

F0 is calculated from short-term autocorrelation and smoothing with Praat [2]. Frames with null F0 values are discarded in the comparison of REP and ALT F0 contours. Also discarded are any frames in which delta-F0 after smoothing is unexpectedly high or low (change of more than 100 Hz in 10 ms). Two methods of pitch comparison are used in this study: trimmed F0 difference and time-normalized F0 difference, but only the time-normalized data are reported below. For trimmed F0, the F0 trajectories of the REP and ALT are compared, where the longer F0 is trimmed to match the length of the shorter F0. For time-normalized F0, the trajectories of REP and ALT are compared, where the shorter F0 is time normalized to match the length of the longer one by using the linear interval interpolation. The mean F0 distance of REP and ALT is then obtained by:

$$\Delta F0 = \frac{\sum_{i,j=1}^{n} (F_0^{(i)} - F_0^{(j)})}{n}$$

Here, i is the ith sample of REP and j is the corresponding sample of ALT, and n is the length of the F0 contours. The F0 difference value is not squared in the equation, because we want to preserve the sign to distinguish cases where REP F0 is scaled higher than ALT from cases which have the opposite scaling relation. We have visually inspected the F0 contour of the REP and ALT sections to be certain that we do not encounter cases where the F0 contours have opposite slopes. Trimming and normalization are two methods we use to guarantee that the F0 contours of REP and ALT that we are comparing have the same length. In this paper, the first F0 values correspond to those of the reparandum (REP) and the second F0 values corresponds to those of the alteration (ALT). Thus, when $\Delta F0 > 0$, REP is higher in F0 than ALT and when $\Delta F0 < 0$, ALT is higher in F0 than REP. Figure 3 shows overlaid time-normalized F0 contours for one REP-ALT pair.

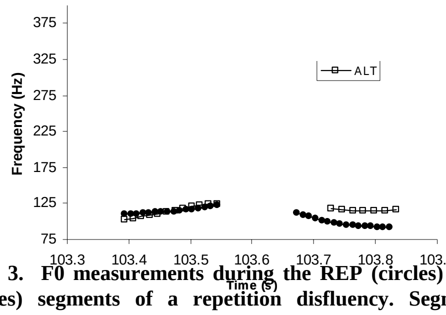

Figure 3. F0 measurements during the REP (circles) and ALT (squares) segments of a repetition disfluency. Segments are aligned using the time normalized F0 difference measurement.

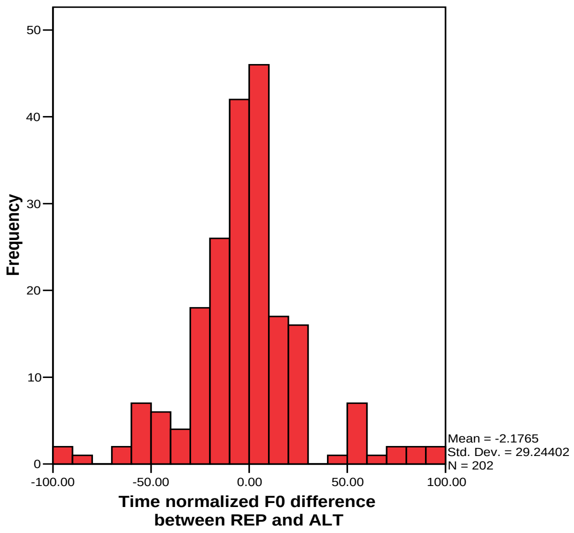

Figure 4. Histogram of time normalized F0 differences between REP and ALT for Hesitation Repetition and Repair Error Correction disfluencies for 202 tokens.

Fig. 4 shows preliminary results that confirm our category labeling results. Fig. 4 is a histogram of the differences in F0 between REP and ALT segments of 202 disfluencies (computed using the time normalized F0 difference measure described in Sec. 4). As suggested by our intonational category labeling results (Table 5), this distribution exhibits three distinct modes: a mode centered at 0Hz average F0 difference, a mode centered at 50Hz average difference (average REP F0 is 50Hz higher than average ALT F0), and a mode centered at -50Hz (average ALT F0 is 50Hz higher). As suggested in Table 5, the mode at 0Hz difference is more than twice as large as the modes at 50Hz and -50Hz. This means that most of the REP-ALT pairs have highly similar F0 contours. The tokens in these three modes of the histogram are not all the same as the tokens in categories hesi-r-same, hesi-r-fp, and hesi-r-exag of Table 5, but there is strong overlap between the modes of the histogram and the labeled categories.

5. Discussion and Conclusion

Labeling of the intonational pattern of repetition disfluencies (Table 5) demonstrated the frequency of four distinct intonational patterns. The most frequent pattern (hesi-r-same, 62 tokens) involved the perceived repetition, in the ALT segment, of the F0 pattern of the REP segment. Three other categories each contained 22-28 tokens: the filled-pause intonational pattern (ALT is produced with a low flat F0 and rapid articulation), the exaggerated intonational pattern (ALT is produced using an exaggerated version of the REP intonation), and the intermediate phrase boundary pattern.

Our quantitative measures of F0 provide suggestive supporting evidence for the parallelism of REP and ALT intonation contours. The prosodic similarity between REP and ALT provides a strong perceptual cue to the listener for the repetition of the lexical item, and may help in the online editing of the disfluency.

6. References

[1] Baron, Don, Elizabeth Shriberg, & Andreas Stolcke. 2002. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. *Proc. ICSLP'02*, Denver, CO, vol. 2, pp. 949-952.

[2] Boersma, Paul & David Weenink. 2005. *Praat: doing phonetics by computer* (version 4.3.04) [Computer Program]. Retrieved March 8, 2005 http://www.praat.org.

[3] Clark, Herbert H. & Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, vol. 84, pp. 73-111.

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

[5] Godfrey, John J., Edward C. Holliman, & Jane McDaniel. 1992. Telephone speech corpus for research and development. *Proc. the International Conference on Acoustics, Speech, and Signal Processing,* March 1992, San Francisco, CA, pp. 517-520.

[6] Heeman, Peter A. & James F. Allen. 1999. Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics,* vol. 25(4), pp. 527-571.

[7] Henry, Sandrine & Berthille Pallaud. 2003. Word fragments and repeats in spontaneous spoken French. *Proc. DiSS'03*, 5-8 September 2003, Goeteborg University, Sweden, pp. 77-80.

[8] Lendvai, Piroska, Antal van den Bosch, & Emile Krahmer. 2003. Memory-based disfluency chunking. *Proc. DiSS'03*, 5-8 September 2003, Goeteborg University, Sweden, pp. 63-66.

[9] Levelt, Willem J. M. 1989. *Speaking. From Intention to Articulation*. Cambridge, MA: MIT Press.

[10] Levelt, William J. M. & Anne Cutler. 1983. Prosodic marking in speech repair. *Journal of Semantics*, vol. 2, pp. 205-217.

[11] Liu, Yang, Elizabeth Shriberg, & Andreas Stolcke. 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. *Proc. Eurospeech*, Geneva, Switzerland, pp. 957-960.

[12] Nakatani, Christine H. & Julia Hirschberg, 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, vol. 95(3), pp. 1603-1616.

[13] Plauché, Madelaine C. & Elizabeth Shriberg. 1999. Data-driven subclassification of disfluent repetitions based on prosodic features. *Proc. International Congress of Phonetic Sciences*, San Francisco, CA, vol. 2, pp. 1513-1516.

[14] Savova, Guergana & Joan Bachenko. 2003. Prosodic features of four types of disfluencies. *Proc. DiSS'03*, Goeteberg University, Sweden, pp. 91-94.

[15] Shriberg, Elizabeth. 1995. Acoustic properties of disfluent repetitions. *Proc. International Congress of Phonetic Sciences*, Stockholm, Sweden, vol. 4, pp. 384-387.

[16] Shriberg, Elizabeth. 1996. Disfluencies in Switchboard. *Proc. ICSLP'96*, 3-6 October 1996, Philadelphia, PA, vol. Addendum, pp. 11-14.

[17] Shriberg, Elizabeth. 2001. To `errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, vol. 31(1), pp.153-164.