DATA-DRIVEN SUBCLASSIFICATION OF DISFLUENT REPETITIONS BASED ON PROSODIC FEATURES

Madelaine C. Plauché*† and Elizabeth E. Shriberg†

*University of California, Berkeley, USA, †SRI International, USA

ABSTRACT

Information about the state and planning of the speaker is obscured in traditional classifications of disfluencies which are generally at the word level. This study delves into the acoustic and prosodic information of *repetitions*, one of the most common disfluencies.   A hierarchical clustering of prosodic features reveals three subsets of repetitions, each reflecting different problems in planning.

1. INTRODUCTION

Spontaneous speech is notorious for its false starts, um's and er's, pauses, fragments, self-repairs, and repeated words. Disfluencies reveal information about the planning problems of the speaker [1, 6] and provide a challenge for speech processing [7].   The descriptive categories used, however, are based on the level of the transcription, which may obscure information in the acoustic signal. For example, Hieke [4] suggests two distinct functions of repetitions (bridging a gap and hesitating) that are confirmed by looking at duration and F0 properties of repetitions [8].

This study continues the subclassification of repetitions using a hierarchical clustering model of prosodic features. We were able to discern three subsets of repetitions based on two commonly repeated function words: 'I' and 'the.' Acoustic and prosodic features such as pause length, relative and absolute word duration, pitch patterns, and non-modal voicing can further classify categories of disfluencies as well as give information about speaker planning.

2. METHOD

In this paper, we will refer to certain parts of the disfluent utterance according to a model adapted from Levelt [1]. The pauses (P1, P2, P3) are optional, but most repetitions involve an original utterance, a first instance of the repeated word (R1), a second instance of the repeated word (R2), and a continuation of the utterance (Fig. 1).

Original Utterance	R1		R2		Continuation
…sticks to (pause)	the	(pause)	the	(pause)	vegetables…
	P1		**P2**		**P3**

Figure 1. Structure of a Repetition.

For this study we are interested in finding functional and descriptive differences between types of repetitions. The data is comprised of repetitions of the words 'I' and 'the,' across 193 different speakers (roughly 36% female and 64% male) to determine generalizations that are not speaker dependent.

2.1. Prosodic Features

Over 800 utterances with a repetition of the word 'I' or 'the' were analyzed from the Switchboard corpus, a spontaneous speech corpus of human-to-human telephone conversations on prescribed topics [3].   Transcribers rated the conversations as highly natural or indistinguishable from natural speech.   A database of utterances containing repetitions was created by hand-labeling acoustic and prosodic features including: pause length, word duration, pitch patterns (values and slopes over words), and presence of non-modal voicing (creak, glottalization, or diplophonia).

2.2. Hierarchical Clustering

As input to our clustering model, the features were first binned and normalized by examining the distribution of values for each feature and specifying 2 or 3 "bins" of values. For example, pause durations were either 0-60, 60-200, or 200+ msec (no pause, short pause, or long pause). See Table 1 for a complete list of the feature bins.

These features were then rank normalized (each feature's value range was from 0 to 1) so that features were rated equally regardless of their distribution across bins. To normalize for inherent differences in word duration ('I' is much shorter than 'the,' in general), the durations were divided by the average duration of the fluent version of that word by speaker. The durations for fluent words were extracted from forced alignments by a speech recognizer, over a large number of fluent tokens (on average: 2,455 tokens per speaker), providing a robust word duration estimate. One of the features extracted is the normalized duration difference of R1 and R2, a ratio of the difference in duration between R1 and R2 divided by the mean duration of fluent instances by word and by speaker. The F0 patterns over the word were grouped into the somewhat simplistic categories of rising, falling, and other, based on straight line approximations [9] of the last slope in the word- stylized F0. Very few tokens fell into the category 'other,' so for further analysis, only rise and fall were used.

The features were then input to an unsupervised hierarchical clustering, which allowed us to determine how tokens group or correlate with each other along the specified features, with no prior hypotheses. We used a simple city-block distance as our metric to evaluate minimal distance in the clustering.

Feature	Bin Variants Before Normalization		
P1 Duration	0-60ms	60-200ms	> 200ms
P2 Duration	0-60ms	60-200ms	> 200ms
P3 Duration	0-60ms	60-200ms	> 200ms
Normalized Duration Difference of R1 and R2: ((R1-R2)/fluent)	Ratio ≤ 0 R2 longer or equal to R1	0 < Ratio ≤ 1 R1 is up to one fluent word length longer than R2	Ratio > 1 R1 is more than one fluent word length longer than R2
R1 F0 Pattern	ends in F0 fall	other (complex or flat)	ends in F0 rise
R2 F0 Pattern	ends in F0 fall	other (complex or flat)	ends in F0 rise
Presence of Creak (R1)	no	yes (anywhere on R1)	
Normalized R1 Duration	< 1.5 times length of fluent version of same word	≥ 1.5 times length of fluent version of same word	
Normalized R2 Duration	< 1.5 times length of fluent version of same word	≥ 1.5 times length of fluent version of same word	

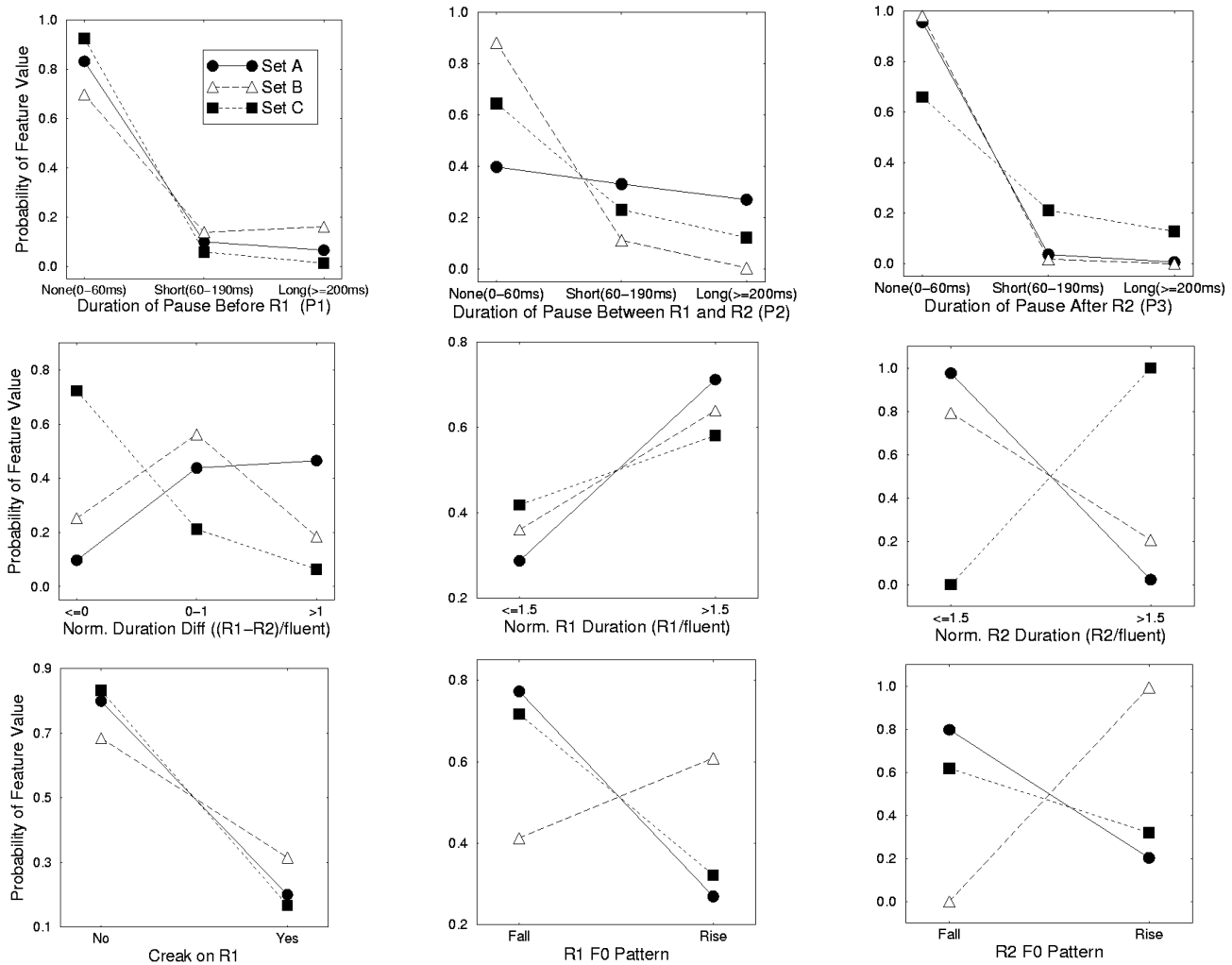Table 1. Feature Bins before Normalization. See Section 1.2. for discussion.

Figure 2. Probability of Feature Values for Sets A, B, and C by Feature. See Section 3. for discussion.

3. RESULTS AND DISCUSSION

Of the 819 original tokens in the analysis, 724 were distributed over three main clusters (Sets A, B, and C). The remaining 95 of the 819 tokens were distributed over more than 32 other clusters, whose distance from A, B, C was larger than that among A, B, C. These outliers represented patterns that occur quite infrequently and therefore were excluded from further analyses. A distribution of the number of tokens by word and by set is provided in Table 2.

Set	"I" Reps		"the" Reps		Total Reps
A	93	31%	206	69%	299
B	138	62%	84	38%	222
C	105	51%	98	49%	203
All Sets	336	46%	388	54%	724

Table 2. Distribution of Tokens by Word and by Set.

For each feature, the probability of feature values by set were plotted along a 0 to 1 axis for each feature (Fig. 2). In Figure 2, the lines connect points from the same set and values of each set sum to one. These graphs convey both relative distribution by feature (along set lines) and absolute distribution of features (along bin columns). For example, the R1 F0 pattern plot (bottom middle) shows that for set A and set C, most R1 tokens end with a fall in F0 and only a few end with a rise in F0. For set B, however, the majority (60%) of R1 tokens end with a rise in F0.

We then wanted to classify the three sets from their prosodic features as well as higher order cues such as position in the phrase, presence of neighboring disfluencies, and relation between original and continuing utterance. From the graphs in Figure 2, we extracted the gross differences of the features for each of the three sets and designed a schematic for the canonical configuration of each set (Fig. 3,4,5). These schematics describe *general* differences extracted from the graphs, but not all tokens of a given set will match the schematics on each feature.
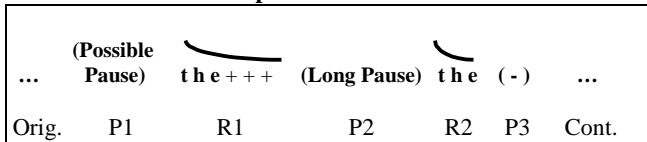
3.1. SetA: Canonical Repetition

Figure 3. Schematic for Set A. Dark lines represent the pitch patterns. '+' = a longer than fluent duration. ' - ' = no pause.

Set A is characterized by a possible P1, an R1 duration that is much longer than a fluent version, a long P2, and an R2 duration that is close to the fluent version with no following pause. The pitch pattern is falling on both R1 and R2. Along these features, Set A corresponds to the *retrospective repeat*, one of two types of repetitions classified by Hieke [4], which was proposed to function as a bridging device to connect a continuation with the original utterance after a break in fluency. Shriberg [8] also found that R1 is much longer than a fluent version and longer than R2. Starting from the assumption that there were two types of repeats, she found she could distinguish between them based on whether or not there was a pause after R2 and before the continuation (P3). As we will see in the following section, both Set A and Set B would fall into this category, as

both are characterized by the lack of a P3. This could explain why retrospective repeats accounted for 90% of all repeats in Shriberg's study.

Presumably the speaker has trouble planning, hesitates or "stalls" either by lengthening R1, pausing, or both, and then resumes the utterance by repeating the head of the syntactic constituent ('I' or 'the') to re-establish fluency for the listener (Ex. 1,2). Support for this claim is found in perceptual experiments that show that some listeners tend not to hear pauses when they are within a constituent [2].

1. … (pau) making all of **the** (pau) **the** family (pau) things work. [SOUND 0679_01.WAV]
2. … we've had a lot of fun. uh (pau) **I I** moved to Dallas about five years ago. [SOUND 0679_02.WAV]

Examples 1 and 2. Tokens from Set A.

Note the relatively high rate of 'the' in Set A (69%). This distribution might be an artifact of the method used to split the files, namely, along turn boundaries. The durations of pauses at the beginnings and ends of files are often misrepresented. Since 'I' occurs commonly in phrase-initial position, where a preceding pause (P1) can be misrepresented, and P1 is a characteristic feature for this set, many 'I' repeats might be missing in the clustering.

We expected to find glottalization at the end of R1 to mark an "abrupt cut-off" at this point. We found very few cases of this, however, in Set A (though see section on Set B), though diplophonia, a creak-like voicing modality which is commonly found on lengthened filled pauses is sometimes found on lengthened cases of R1. This suggests that the speaker is using a similar strategy in the production of R1 as in filled pauses.

We draw a distinction between a state in which the speaker has not yet committed to following material and that in which the speaker has made a covert repair once a problem was detected. Set A matches what we would expect a signal to look like in the former case. The utterance shows problems up until R1 (sometimes not until R1), but R2 is simply a fluent continuation of the first part of the utterance, with no evidence of further trouble in planning.
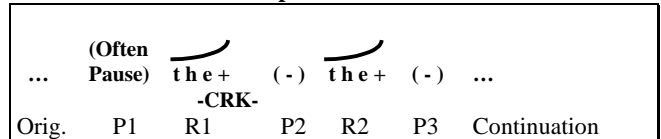
3.2. Set B: Covert Self-Repairs

Figure 4. Schematic for Set B. Dark lines represent the pitch pattern. '-CRK-' = presence of creaky voice or glottalization. '+' = longer than fluent duration. ' – ' = no pause.

Characteristic of Set B is the presence of P1 but no other pauses, a slightly long R1 with rising pitch and sometimes glottalization, and a slightly long R2 also with rising pitch. This does not match the hesitation-type disfluency found in Set A; instead it most closely corresponds to repetitions that are *(covert) self-repairs* [6]. Whereas in Set A (Ex. 1,2), the R2 is often simply a continuation of the utterance with one word repeated, in

Set B, R2 usually marks the beginning of a new utterance or a corrected version of the previous one; it is not a mere continuation (Ex. 3,4).

3. I I think (pau) **the the** (pau) thing is though, I I guess…
   [SOUND 0679_03.WAV]
4. (pau) I think I think (pau) I think (pau) you know for myself (pau) **I I** (pau) see that as probably the …
   [SOUND 0679_04.WAV]

Examples 3 and 4. Tokens from Set B.

The pause preceding R1 and the glottalization on R1 suggest that the speaker detects a problem (perhaps an inconsistency with what was previously said or a problem later on in the utterance) during the production of R1. The rise in pitch that is repeated over R2 corroborates this, as it shows the speaker is resetting the signal or emphasizing what is to follow relative to what was just said. There is often evidence of problems before R1 (pauses, other disfluencies), but the speaker stops herself at R1, frequently with an abrupt cut-off. Jasperson [5] claims that this type of cut-off enables the soonest possible resumptions, which would also explain why it is often found in the self-repair set of repetitions.

If Set B is a cluster of self-repairs, this could also explain the majority of tokens of 'I' repeats (62%), as they are more likely to head a heavy constituent, which is found to be correlated with planning difficulty for the speaker [1]. This is confirmed by other information in the signal that indicate major problems and stalling; phrases such as 'I mean,' 'you know,' and 'the thing about it is' were very commonly found in the original utterances of this set.

Whereas Set A exhibited cases of the speaker not yet having committed to following material, Set B shows evidence that the speaker has detected a problem. The speaker detects this problem at R1 and stops herself, then continues once she has (re)planned this utterance. The detected problem can be later in the utterance, in the case of heavy constituents (Ex. 4), or it might be in the original utterance, in which case R2 is a repair of the previous utterance, apparent in the lexical structure (Ex. 3).

3.3. Set C: Stalling Repetition

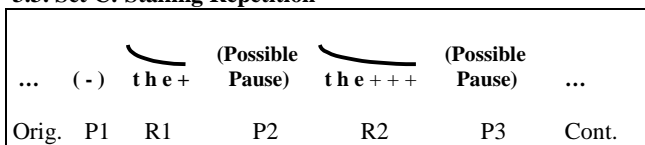| …	(-)	t h e +	(Possible Pause)	t h e + + +	(Possible Pause)	…
Orig.	P1	R1	P2	R2	P3	Cont.

Figure 5. Schematic for Set C. Dark lines represent the pitch patterns. '+' = longer than fluent duration. ' – ' = no pause.

The canonical Set C token has no preceding pause, a slightly longer R1 that is dropping in pitch, a possible value for P2 and P3, and an R2 that is much longer than a fluent version. Descriptively, Set C looks very much like the inverse of Set A. In these tokens, the speaker has problems during and after the production of R2. This is seen by the lengthening of R2 and the possible presence of a pause following R2 (recall that there is no pause after R2 in Set A or B). This is also the only case where there is no pause before R1, suggesting that the speaker does not

run into problems until after she has begun production of R1. More evidence that the speaker is continuing to have trouble after R2 is the presence of other disfluencies and hedges immediately after the repetition (Ex. 5,6). This was not the case in Set A, where other disfluencies, if present, were before R1 and where R2 and the continuation were fluent. Here the speaker is perfectly fluent up until R1, though trouble in planning might start before.

5. …(pau) don't have **the** (pau) **the** (pau) special tools or (pau)… [SOUND 0679_05.WAV]
6. …(pau) no **I I** think uh (pau) I uh. [SOUND 0679_06.WAV]

Examples 5 and 6. Tokens for Set C.

This case matches the prosodic features of Hieke's [4] *prospective repeats* (long P3 pause, lengthened R2). These repeats may serve a stalling function, to hold the floor during major problems in planning.

4. CONCLUSION
Three subsets of repetitions can be determined from a hierarchical clustering of acoustic features. Incorporating prosodic information, which is obscured at the text level, in an analysis of disfluencies is crucial in suggesting how repeats relate to speaker states and strategies.

REFERENCES
[1] Clark, H. and T. Wasow. 1998. Repeating Words in Spontaneous Speech. *Cognitive Psychology*, 37, 201-242.
[2] Duez, D. 1985. Perception of silent pauses in continuous speech. *Language and speech*, 28(4), 377-389.
[3] Godfrey, J., E. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. *IEEE Proc. of the IEEE Conf. on Acoustics, Speech, and Signal Processing, San Francisco*, 1, 517-520.
[4] Hieke, A.E. 1981. A Content-Processing View of Hesitation Phenomena. *Language and Speech*, 24 (2), 147-160.
[5] Jasperson, R. 1998. *Repair after cut-off: Explorations in the grammar of focused repair of the turn-constructional unit-so-far.* Unpublished Ph.D. Thesis, University of Colorado.
[6] Levelt, W. 1983. Monitoring and self-repair in speech. *Cognition*, 14, 41-104.
[7] Nakatani, C. and J. Hirschberg. 1993. A Speech-First Model for Repair Detection and Correction. In *Proc. of the Assoc. for Computational Linguistic*s, *Ohio State University,* 46-53.
[8] Shriberg, E. E. 1995. Acoustic properties of disfluent repetitions. *Proc. Intl. Congress of Phonetic Sciences, Stockholm,*,4, 384-387.
[9] Sonmez K., E. E. Shriberg, L. Heck, and M. Weintraub. 1998. Modeling Dynamic Prosodic Variation for Speaker Verification. *Proc. of the Intl. Conference on Spoken Language Processing,* 7, 3189-3192.