10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

The Prosody of Fluent Repetitions in Spontaneous Speech

Hong Zhang[1]

[1]University of Pennsylvania

zhangho@sas.upenn.edu

Abstract

Repetitions in spontaneous speech have traditionally been regarded as disfluencies that function either as a form of hesitation or a tool to maintain continuity after a break. However viewing repetitions as disfluencies leaves many repetitive phenomena in spontaneous speech unexplained. We show that a common form of repetitions, which is described as the rapid repeating of single syllable function words two or more times, can be distinguished from disfluent repetitions through simple prosodic characterizations such as silent pause durations after repetition and F0 contours and the duration of the repeated phrases. The duration and F0 features of these fluent repetitions failed to support the idea that these repetitions involve the same cognitive process as repair. We propose that such fluent repetitions are linked at least to the generation and execution of motor plans, with possible extensions to higher level planning such as at the level of morphosyntax or even semantics.

Index Terms: repetitions, speech disfluencies, speech planning

1. Introduction

Disfluencies in spontaneous speech, albeit the different forms they may take, are often thought to arise from hesitation or the need for repair [1]. Both hesitation and repair are linked to problems during speech planning. It has been argued that details of the planning problem can be identified through both textual and acoustic or prosodic features of surface realizations of disfluencies. For example, two forms of filled pauses, *um* and *uh*, have been claimed to mark major and minor delays during the delivery of speech [2]. Extensive analyses of the acoustic properties of repetition and repair disfluencies [3, 4, 5] have found support for the subclassifications proposed in psycholinguistic literature [6, 7, 8].

Given the close connection between surface disfluencies and the underlying cognitive process of speech planning, researchers often either implicitly or explicitly link their empirical descriptions to psycholinguistic models of speech production, most notably the Nijmegen tradition which involves some speech monitoring mechanism [8, 9]. For repetitions in particular, accounts often follow the bipartite distinction between a sign of hesitation, that is, buying time for message formulation, and a special case of repair [6, 1]. Combining observations from the textual properties of repetitions that repeated words mostly consist of function words, [10] propose a Commit and Restore model. Although this model deviates from others in terms of avoiding bringing up speech monitors, the basic assumption that repetitions are forms of disfluencies is maintained.

The focus of our current study is a kind of repetition which can be characterized as rapid fluent repetitions of mostly single syllable function words at the beginning of some syntactic or prosodic phrase in utterances that are otherwise fluent and well structured. Some example transcriptions of such repetitions, extracted from SCOTUS 2001 corpus [11], are shown in figure 1. Models based on the disfluency assumption of repetitions are unable to offer an adequate account in explaining these fluent repetitions primarily due to the need for time for planning or replanning, which will be shown largely absent. We will also present evidence that fluent repetitions on average do not modify the duration of repeated phrases to an extent sufficient for marking a delay due to planning needs. F0 reset between the second and first repeat is also not observed in the fluent type. Therefore it is unlikely that replanning is involved. We propose that alternative accounts from the perspective of speech motor planning and execution could offer more plausible explanations to these fluent repetitions.

but ***it's it's*** really the disability that we're focusing …

i i didn't find it perhaps you could point me to …

though in this case ***i'm i'm*** sure you would contest …

Figure 1: *Example transcriptions of fluent repetitions.*

2. Method

In this study, we adopt a model from Levelt [7] and Plauché and Shriberg [3] to refer to different parts of a repetition. To simplify the discussion even further, we only consider repetitions that involve repeating same word or phrase twice. As illustrated in figure 2, the segments of interest include pauses before (P1), between (P2) and after (P3) the two repeats, as well as the first (R1) and second (R2) occurrence of the repeated phrase.

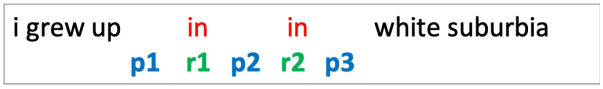

Figure 2: *The structure of repetitions.*

2.1. Corpus, sub-sampling, and data annotation

A subsample from the Fisher corpus [12], a large collection of English spontaneous telephone conversations, was used for the current study. Speech from 200 native speakers of American English was randomly selected to construct the analysis sample. Repetition instances were first automatically identified using a method adapted from the Suffix Tree algorithm. Then for each speaker, up to 5 random instances of repetitions were selected after manual examination and correction. The selected examples were further classified into three types of repetitions: *fluent*, *delayed*, and *disfluent*, based on the criteria summarized in table 1. A total of 743 repetition examples were collected and classified for the analysis.

The threshold of 150ms silence between repeats was selected to acknowledge both the perceptual effect of a silent

10.21437/SpeechProsody.2020-155

Table 1: *Criteria used for repetition subclassification*

Repetition type	**Criteria**
Fluent	$P2 < 150ms$ or no perceptible pause between repeats; No other disfluencies in the same utterance.
Delayed	$P2 \geq 150ms$ or perceptible pause between repeats; No other disfluencies in the same utterance;
Disfluent	The repetition is part of other disfluencies in the same utterance; The repetition is immediately following, or followed by other disfluencies.

pause and the expected time that at least motor replanning would take [13, 9]. Therefore the *delayed* type can be regarded as representing repetitions potentially caused by minor breaks or brief hesitations. This is in contrast to repetitions that are co-occurring with other disfluency phenomena, which are supposedly linked to disruptions involving larger discourse structure.

2.2. Acoustic measurements

Duration measurements were based on forced alignment result using the HMM-based Penn forced aligner. F0 measurements were obtained using a pitch tracker which implements the autocorrelation method described in [14]. Raw F0 measurements were smoothed via quadratic interpolation. To compare the F0 contours between R2 and R1, the smoothed F0 curves were projected onto an orthogonal functional basis defined by the first five Chebyshev Polynomials of the First Kind:

$$T_n(z) = \frac{1}{4\pi i} \oint \frac{(1-t^2)t^{-n-1}}{(1-2tz+t^2)} dt \qquad (1)$$

A plot of the basis functions is shown in figure 3. The coefficients attached to the basis functions after linear projection can get natural interpretations related to the overall F0 height, slope and higher order curvature of the contours under comparison.

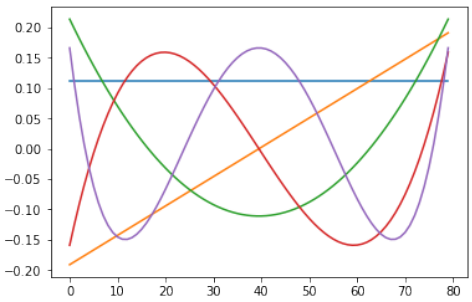

Figure 3: *Plot of the shape of the basis functions.*

3. Results

The distribution of repetitions by type is summarized in table 2. The table shows that the majority of labeled repetitions are either fluent or containing some delay between repeats in otherwise fluent utterances. In addition, most of the repeats are composed of single syllable function words. This distributional property is consistent with previous studies using Switchboard [10]. Since our classification criteria subjectively define the silent pause duration between repeats by type (P2), and P1 is often representing the silence of a long break by design, their duration distributions are less informative compared to P3, which we didn't control during subclassification. Therefore the duration of P3, duration change from R1 to R2 and the F0 contours of R1 and R2 can be used to test the hypothesis that the three types of repetitions are subject to different disruptions during production. We would predict that *fluent* type is at least not directly linked to hesitation or speech error, while *delayed* and *disfluent* types are related to minor and/or major (discourse structural) hesitation or repair.

Table 2: *The distribution of repetition type and repeated phrase type*

Repetition type	**Count/%**	**Single Syl. %**
Fluent	390/52.5	87.2
Delayed	164/22.1	82.3
Disfluent	189/25.4	86.2

3.1. Pause duration after repetition

If the hypothesis holds, it can be expected that fluent repetitions will on average have very short delays in P3, while disfluent repetitions will have longer P3. Figure 4 plots P3 duration in three types of repetitions.

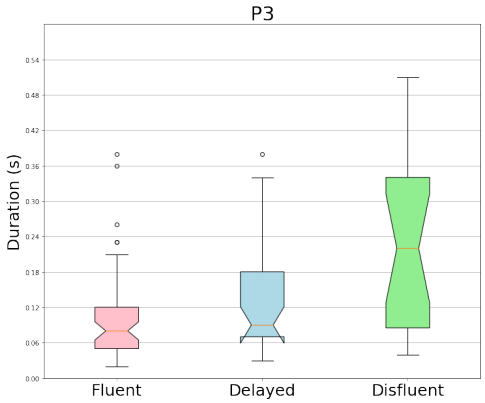

Figure 4: *Silent pause duration following repetitions.*

As expected, there is an increase in the overall P3 duration from *fluent* to *disfluent* repetitions, with the *delayed* somewhere in between. Most noticeably, P3 durations are almost always shorter than 200ms in the *fluent* type, and the 4th quartile in the *delayed* type is also only 180ms. These short pauses are not sufficient for making an alternative speech plan or even motor plan. On the other hand, the median duration of P3 in *disfluent* repetitions is over 200ms, suggesting the existence of some major planning between repetition and the following delivery of speech. Thus observations from P3 duration distribution support our hypothesis.

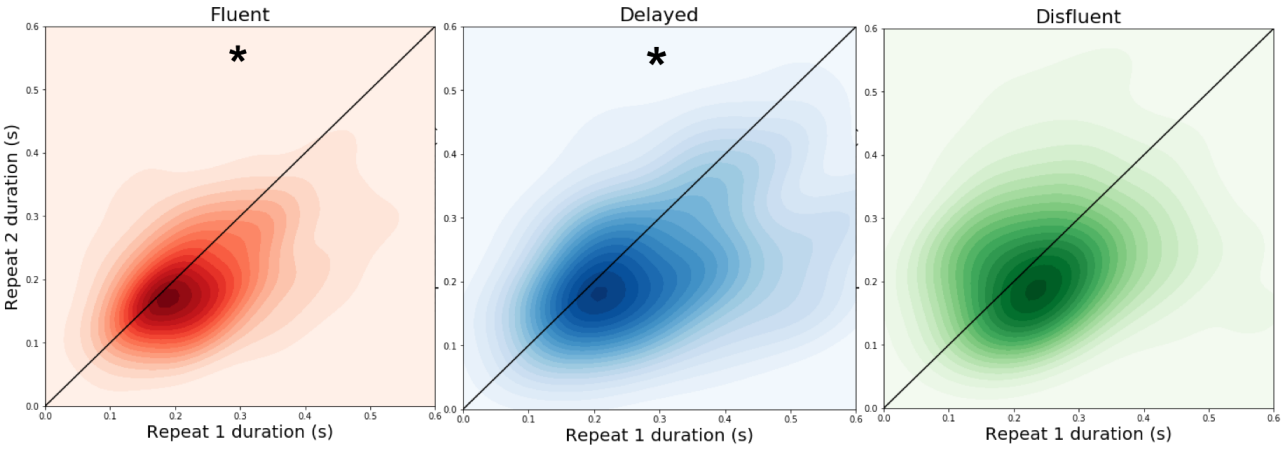

Figure 5: *Duration relationship between R1 and R2.*

3.2. Duration of the repeats

Duration relationships between R1 and R2 can offer another piece of evidence to test whether repetitions involve replanning. Under Hieke's [6] dichotomy of retrospective and prospective repetition, longer R1 should be related to some latent repair process right before articulation, while longer R2 indicates hesitation. This distinction has been supported by [5] using Switchboard. Here we would like to ask if this dichotomy can be observed from our proposed repetition types.

Figure 5 plots the density distribution of duration of R1 against R2 on in three types. It can be observed that in the *fluent* type, the distribution is tightly clustered along the equal duration line, with small but significant difference between R1 and R2 ($p < 0.001$), where R1 is longer. More density is distributed towards the lower right corner of the graph for the *delayed* type, which indicates longer R1 than R2. This difference is also significant ($p < 0.001$). However, in the *disfluent* type, the distribution spreads to both upper left and lower right corner, and here the difference is not significant ($p = 0.2$). This suggests that although for each R1-R2 pair, the duration is more likely to show larger difference compared to the fluent repetitions, there isn't a predominant trend: both hesitation or repair can be present in this type of repetition.

The distribution in the *fluent* type is not likely to be associated with the retrospective repetition as proposed in [15]. Comparing this graph with what have been reported in [5], where the duration difference is often found to be greater than 0.1s, the duration difference between R1 and R2 in our sample is quite small, such that the extra time is not sufficient for restructuring or planning for the coming repair. Thus the slight but consistent reduction in R2 duration can be better explained as shortening of R2 due to reduction caused by repeating the same word twice, rather than lengthening of R1. In addition, two ridges can be observed from the *delayed* type: one almost aligns with the diagonal, and the other is off diagonal in the lower right corner. This may suggest a further split within the delayed category: some repetitions are probably indeed caused by retrospective repair, while others are more similar to the *fluent* type.

3.3. F0 contour of repeats

The next test of our hypothesis that the three repetition types are related to different underlying processes can be found through comparisons between F0 contours in R1 and R2. In the functional space defined by the first 5 Chebyshev Polynomials of the First Kind, as described above, coefficients attached to polynomials can get intuitive interpretations related to the shape of F0 contours. The first coefficient can be interpreted as representing the overall F0 height, and the second coefficient can be understood as representing the average slope. Coefficients of the third and higher order functions are then representing the more complex curvature properties. Pairwise Wilcoxin signed rank tests were run for each repetition type and in all five dimensions, but significant difference was only found in the first dimension. Therefore the change in F0 contour between R1 and R2, if there is any, is mainly present in terms of the overall F0 height, but not the slope or other aspect of F0 curvature.

Figure 6 plots the coefficients of R1 in the first dimension against R2 across three types. The dotted line in each graph shows the estimated linear regression line. In both *fluent* and *delayed* types, significant difference was found in the pairwise comparison between R1 and R2 F0 contours, suggesting that the overall F0 height is lower in R2 compared to R1. The estimated regression line also clearly shows an off-diagonal trend in both types. However, Significant difference was not found in the *disfluent* type, and the regression line almost overlaps with the equilibrium line. Therefore the overall F0 height is about the same in R1 and R2 in this condition.

Lower F0 in R2 can be interpreted as relating to pitch lowering in speech production. The lack of pitch lowering in the *disfluent* type indicates the potential existence of F0 reset in the second repeat within a repetition. Therefore it is probable that the second articulation of a same form is the result of replanning. On the contrary, the descending in F0 height in the other two types suggests the lack of reformulation of an utterance plan, at least at the level of some prosodic phrase.

4. Discussion

The results reported so far all point to a clear prosodic distinction among the three types repetitions that we have proposed. In particular, all three measures: the silence duration following the repetition, the duration and F0 height difference between two repeats, all suggest that our *fluent* type is distinctive from the other two types, and probably indeed fluent. However, the *delayed* type should be viewed as a mixture of more fluent repetitions and those that are caused by minor repair, in the sense of Hieke's retrospective repetition. The distribution of the three

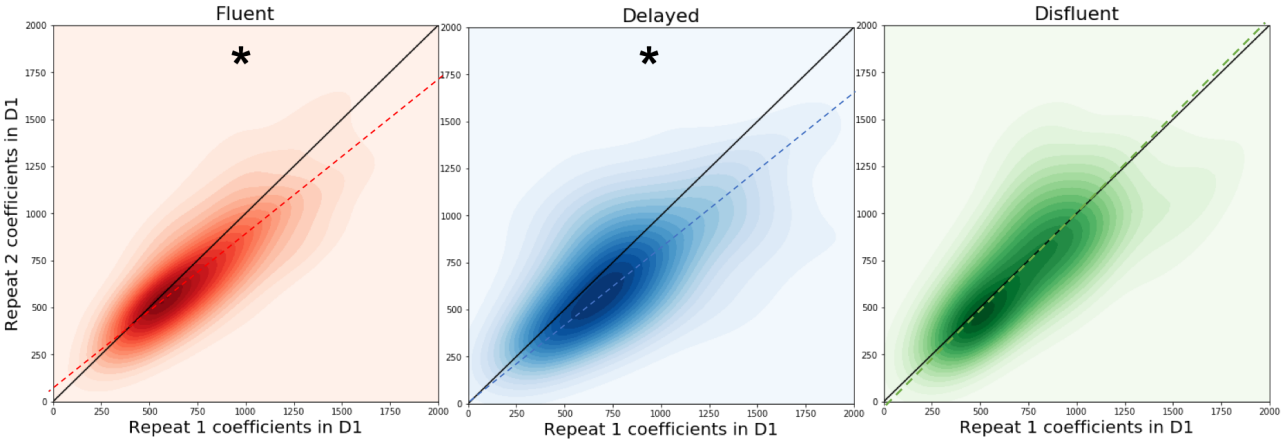

Figure 6: *F0 difference in contour between R1 and R2.*

types in our sample shows that most of repetition instances are from the *fluent* or *delayed* type, with fluent repetitions occupy the largest share. Thus we could first argue that most of the repetitions found in spontaneous conversations are likely to be fluent. Characterizing all repetitions indiscriminately as disfluent is therefore not entirely appropriate.

If the observations from our small but representative sample hold at larger scale, we are then facing a problem of not having a theory to account for the majority of repetition phenomena common in spontaneous speech. First off, since we just made the case that most repetitions are in fact quite fluent, theories stemming from speech disfluencies are not applicable, since the existence of hesitation or repair has already put into question. Speech monitoring models, such as those derived from the Nijmegen model of speech production [8] will also face the apparent problem of not having sufficient time for the feedback loop to finish its duty. It has generally been assumed that the temporal commitment from processing feedbacks from some monitor to reformulate an utterance plan is at least 200 to 250ms, which is much longer than our classification criterion for fluent repetition, and is also longer than the silence duration immediately following repetitions. Therefore it is not likely that fluent repetitions would trigger speech planning all the way up to conceptualization.

An explanation from motor planning and execution perspective can be tenable. Under this view, the production of fluent repetitions is fundamentally a motor control problem. Details of some popular speech motor control models are explained in [16, 17, 15]. A plausible scenario could be the following: the feedback loop in the motor control system mistakenly detects an error signal in the output from the forward model. This false positive then leads to the execution of an existing plan twice. This process is much faster than the one described previously, with an estimate somewhere between 60 and 120ms [13], and is consistent with the durations of P2 and P3 we observed from our sample. Unfortunately, relying solely on such a view may still only lead to an incomplete theory: if fluent repetitions are the result of motor control problem, it would not be unreasonable to expect similar repetitions occurring with phonologically more complex words, or smaller phonological unit such as the first or first several syllable of a word. However, this doesn't seem to be the case. As table 2 shows, the majority of repetitions are single syllable function words. Thus the role of word boundary and word type should be jointly considered.

With the considerations sketched above, we propose that to achieve a plausible explanation to these fluent repetitions, the problem should first be formulated as a motor control problem during the motor planning and execution phase of speech production. However, higher level processes are expected to affect how likely a duplication of motor command is for a given lexical input. As proposed in [15], it is not unlikely that some motor control mechanism can be found in higher levels of the planning process, such as at the level related to morphosyntax or even semantics. Unfortunately, with our limited data, it is not possible to postulate further speculations on whether the duplication of command happens during lexical selection, morphosyntactic structuring or during the transmission of signals between levels of processing. A more comprehensive overview of similar phenomena across languages and populations (i.e., people with known neurological disorders or under the influence of substances that affect language related cognitive function) could offer more insights toward a plausible theory of fluent repetitions.

5. Conclusions

In this study, we offered detailed descriptions of three prosodic properties of repetitions, a common phenomenon in spontaneous speech. Our analyses based on a random sample of telephone conversations from the Fisher corpus found strong evidence that most repetitions are in fact not disfluent, in the sense of being the result of hesitation or the need for repair. Acknowledging the fact that repetitions are more likely to be fluent than disfluent poses a problem for speech production models. We reviewed limitations of potential explanations from the view of both speech monitoring models and motor control, and discussed the possibility of combining higher level planning mechanism with motor planning and execution as a plausible direction towards a theory of fluent repetitions. Future advances toward a fuller theory should draw evidence from more diverse languages and speaker populations.

6. Acknowledgements

We would thank LDC for providing the data, and audience at Penn Commonground for valuable feedback on this project.

7. References

[1] R. J. Lickley, "Fluency and disfluency," *The handbook of speech production*, p. 445, 2015.

[2] H. H. Clark and J. E. F. Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[3] M. Plauché and E. Shriberg, "Data-driven subclassification of disfluent repetitions based on prosodic features," in *Proc. International Congress of Phonetic Sciences*, vol. 2. Citeseer, 1999, pp. 1513–1516.

[4] C. H. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1603–1616, 1994.

[5] E. Shriberg, "Acoustic properties of disfluent repetitions," in *Proceedings of the international congress of phonetic sciences*, vol. 4, 1995, pp. 384–387.

[6] A. E. Hieke, "A content-processing view of hesitation phenomena," *Language and Speech*, vol. 24, no. 2, pp. 147–160, 1981.

[7] W. J. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.

[8] W. J. Levelt and M. Speaking, "From intention to articulation," *Cambridge, MA: The MITPress*, 1989.

[9] A. Postma, "Detection of errors during speech production: A review of speech monitoring models," *Cognition*, vol. 77, no. 2, pp. 97–132, 2000.

[10] H. H. Clark and T. Wasow, "Repeating words in spontaneous speech," *Cognitive psychology*, vol. 37, no. 3, pp. 201–242, 1998.

[11] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[12] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.

[13] O. Civier, S. M. Tasko, and F. H. Guenther, "Overreliance on auditory feedback may lead to sound/syllable repetitions: simulations of stuttering and fluency-inducing conditions with a neural model of speech production," *Journal of fluency disorders*, vol. 35, no. 3, pp. 246–279, 2010.

[14] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[15] G. Hickok, "Computational neuroanatomy of speech production," *Nature Reviews Neuroscience*, vol. 13, no. 2, p. 135, 2012.

[16] F. H. Guenther, "Cortical interactions underlying the production of speech sounds," *Journal of communication disorders*, vol. 39, no. 5, pp. 350–365, 2006.

[17] J. W. Bohland, D. Bullock, and F. H. Guenther, "Neural representations and mechanisms for the performance of simple speech sequences," *Journal of cognitive neuroscience*, vol. 22, no. 7, pp. 1504–1529, 2010.