12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)
4-5 September 2025, Lisbon, Portugal

The relationships between speech rate, disfluencies and reduction phenomena

Emanuela Gallo[1], *Giuseppe Magistro*[1,2], *Francesca Garofalo*[1], *Loredana Schettino*[3]

[1]University of Naples "Federico II", Italy
[2]Ghent University, Belgium
[3]Free University of Bozen-Bolzano, Italy

emanuela.gallo3@unina.it, giuseppe.magistro@unina.it, loredana.schettino@unibz.it

Abstract

Disfluencies and phonetic reduction are natural features of spoken communication. It has been shown that several linguistic factors influence their production, with speech rate considered as a strong indicator of phonetic variation and also related to the occurrence of various types of disfluencies. In this study, we explore the relationship between speech rate, disfluency, and phonetic reduction in spontaneous Italian speech. As expected, our results indicate that higher speech rates generally correlate with fewer disfluency occurrences and with more frequent phonetic reductions. However, these phenomena are subject to limitations: after a certain threshold, further increases in speech rate no longer significantly reduce disfluencies, and at the same time, the rate at which phonetic reductions increase slows down at higher speech rates.[1]

Index Terms: reductions, speech rate, methodology

1. Introduction

The online processes of speech production often result in utterances that are characterized by the occurrence of disfluencies and phonetic reduction. The former refers to heterogeneous phenomena that delay or interrupt the speech flow, such as pauses, fillers, repetitions, and corrections (see [1] for an overview). The latter refers to phonetic segments that are underspecified in the acoustic signal, ranging gradually from vowel centralization or consonant lenition to the deletion of entire segments or syllables [2]. Both phenomena may be partially related to speech rate (henceforth SR), defined as an overall measure that includes articulation rate and pause time [3]. Indeed, SR is considered a robust predictor of phonetic variation [4], and it has been shown to affect the occurrence of different types of disfluencies, with faster speech generally leading to more repetitions, though not more fillers [5]. Furthermore, the type of disfluency also plays a role in modulating phonetic variation. For instance, [6] investigated the role of disfluencies as a contextual variable in word form variation in English conversation, focusing on monosyllabic functional words. While such words frequently undergo reduction due to high frequency and high contextual predictability, the study found that when adjacent to disfluencies — particularly those with a prospective function — they tend to be longer or have a fuller form, which reflects a speaker's difficulty in planning upcoming content. On the other hand, [7] analyzed the relationship between informational redundancy and pronunciation variants in self-repairs in Dutch speech by measuring SR and segmental reduction in both the reparandum and the repair stretch. Despite the potential informational load of repair stretches, they were produced at higher SRs and with greater segmental reduction than reparanda, suggesting that speakers prioritize resuming discourse fluency potentially as a face-saving strategy to move past the error and its correction as quickly as possible. Based on this framework, the main aim of this study is to provide a preliminary exploration of the relationships between SR variations, disfluencies, and phonetic variation phenomena in spontaneous Italian speech. Specifically, we address the following research questions:
Q1: How are variations in SR related to the occurrence of disfluencies?
Q2: How is SR related to phonetic reduction?
Despite the fact that surveying these RQs may yield predictable results, we will assess these research questions with a complex quantitative methodology to detect whether these relationships are always consistent throughout different levels of SR.

2. Methodology

2.1. Corpus

The analysis was conducted on the "Modokit-FROG" corpus[2], consisting of audio-visual recordings of 10 Italian university students, recorded by a Zoom H5 microphone and a professional camera Nikon D7000. We have taken a sub-set of this original corpus, comprising data from 4 speakers from Teramo. The limited sample of speakers is justified by the detailed phonetic analysis required, which involves a comprehensive multitier manual annotation with the aim of capturing various types of disfluencies and phonetic reductions. Furthermore, this paper is intended to establish methodological foundations and identify patterns that warrant further investigation with an expanded speaker sample in future research. Following the Modokit protocol [8], participants were divided into two groups, each tasked with narrating the picture book "Frog, ¿where are you?" by Mercer Mayer (1969) through both oral and written productions. The book narrates the story of a young boy and his dog who are looking for a frog that had escaped. The oral rendition consisted in a speaker who actively participated by telling the story to a silent interlocutor, who was unaware of the story's content. For the written rendition, participants were asked to write a description of what they observed in each image. The first group of speakers (PS01 and PS05) completed the oral nar-

[1]This article is the result of the collaboration among the authors. However, for academic purpose only, Emanuela Gallo is responsible for writing the first draft, data curation and annotation. Giuseppe Magistro is responsible for writing the first draft, data analysis, supervision. Loredana Schettino is responsible for writing the first draft, supervision, data curation, data analysis. Francesca Garofalo is responsible for data annotation.

[2]Sarro, G. (2022). *The many ways to search for an Italian frog. The manner encoding in an Italian corpus collected with Modokit*. Master thesis: Università degli Studi dell'Aquila.

10.21437/DiSS.2025-2

ration before the written task, whereas the second group (SP09 and SP10) started with the written narration, followed by the oral one.

2.2. Reduction Quantification

Reduction phenomena have been quantified (see [9]) by manually editing syllable segmentation in Praat [10] on the expected (phonologic) and observed (phonetic) levels, following the Sonority Sequencing and Maximum Onset Principles. The alignment between the sequence on both levels was evaluated using SCLITE, which is a tool included in the Speech Recognition Scoring Toolkit (SCTK) provided by the American National Institute of Standards and Technology (NIST). The evaluation output reports the following cases:

- "Deletion", a phonetic syllable is found in place of two phonological ones;
- "Substitutions", a phonetic syllable is different from the corresponding phonological one.

2.3. Disfluency Annotation

Disfluency phenomena have been annotated with ELAN [11] according to a multilevel annotation scheme that provides information at formal and functional levels [12]. The first level concerns the macro-structures of the disfluencies in which several regions are delineated, namely, the region to be repaired (RP), the repaired one (RS), and the one in which the delay occurs (IM). The second level focuses on the micro-structure of disfluency, providing detailed information about the specific type of disfluency phenomenon. At the third level, each item is assigned its macro-function, categorized as either Backward-Looking (BLD) or Forward-Looking disfluency (FLD) [13]. On a fourth level, Forward-Looking items are associated with more specific possible functions based on their context of occurrence:

- Word Searching (WS) for items that show difficulties retrieving the target word.
- Structuring (STR) for items involved in structuring discourse at a syntactic and informative level.
- Focusing (FOC) for items preceding semantically relevant elements.
- Hesitative (HES) is the label assigned to items not involved in any of the other specified cases and just associated to a generic speech planning function.

2.4. Prosodic Annotation

To test the aforementioned hypotheses in operational units, the corpus was divided into intonational phrases (in the senses of [14, 15, 16]). In the annotation of the boundaries between intonational phrases, a series of acoustic cues were taken into account namely pitch declinations after the realization of a fully-fledged contour, pitch resets, final lengthenings, and severe interruptions of the phonation flow (breathing pauses or inhalations). Aware of the discretionary nature of the task, we performed an inter-annotator agreement task (as advocated by [17] for the annotation of intonational phrases). Based on the criteria mentioned above, two trained phoneticians annotated separately $\sim$3 minutes of the corpus on Praat TextGrids. In comparing both annotations, a tolerance of 0.05 seconds was observed to account for minimal fluctuations in the boundary placement. Cohen's κ coefficient showed a good agreement between the two annotators (κ = 0.62, Z= 11.17 , SE = 0.06, p $<$ .05). Following this validation step, the remaining data was annotated by one of the authors using the same guidelines.

2.5. Analysis

To investigate how SR is related to the occurrence of disfluency and reduction phenomena in the observed Italian data, SR was defined as the number of expected syl/s and computed per intonational unit. Note that we considered SR where phonation started, i.e. by excluding long breaks or pauses between intonational units. SR values were then normalized per speaker to account for speaker specificity. To answer Q1 and Q2, we fitted different models in R. Given that we are using the number of disfluencies and reductions as dependent variables, we fitted Poisson regressions. Note, however, that there were several intonational units with no disfluencies, leading to potential issues in the estimation of coefficients for Poisson regressions. To circumvent this limitation, we initially fitted zero-inflated Poisson regressions. However, upon checking the statistical assumptions of the models, the relationships between variables (number of disfluencies, reductions and SR) emerged as non-linear, leading to a poor fit. Hence, we have decided to use Generalised Additive Mixed Models (GAMMs) with a Zero-inflated Poisson family distribution (ziP) by deploying the R library *mgcv* [18]. GAMMs provide a more stable methodology for capturing non-linear effects.

3. Results

The dataset consists of approximately 35 minutes of recordings. Table 1 outlines the duration and the occurrences of disfluency and reduction phenomena for each speaker.

Speaker	Dur (m)	Dis (n)	Red (n)	SR $\pm$ SD (syl/s)
PS01	10.26	378	314	6.3 $\pm$ 1.8
PS05	10.05	216	175	4.6 $\pm$ 1.3
SP09	06.59	191	81	4.9 $\pm$ 1.3
SP10	06.27	155	107	4.9 $\pm$ 1.1

Table 1: *Duration (Dur), number of disfluencies (Dis), reduction phenomena (Red), average Speech Rate (SR) and SR standard deviation (SD) for each speaker in the corpus*

Among the total number of disfluencies, N = 940, only 99 fell within the category of Backward-looking disfluencies (BLD), showing an imbalance in the data. Hence, we have decided to consider the number of disfluencies without splitting in subcategories.

3.1. Number of disfluences

The first GAMM examined the relationship between the number of disfluencies in each intonational phrase and SR. The SR of each intonational unit was normalized for each speaker to offset individual differences. Finally, we also included a random intercept for the speaker. To sum up, we fitted a model with the formula in 1.

$$gam(no\_disfl \sim s(SR\_Norm) + s(Speaker, bs = re) \quad (1)$$

The key statistical assumptions for the model were checked with the function *gam_check* and they were met satisfactorily. The effective degrees of freedom (edf), corresponding to 4 for SR, confirm the non-linearity of the effect. The results of the model are reported in Table 2.

Parametric	coefficients			
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.51	0.08	6.19	< .001
Approximate	significance	of smooths		
	edf	Ref.df	Chi.sq	p-value
s(SR_Norm)	4.00	4.97	44.49	< .001
s(Speaker)	2.57	3.00	22.51	< .001

Table 2: *GAMM model with number of disfluencies*

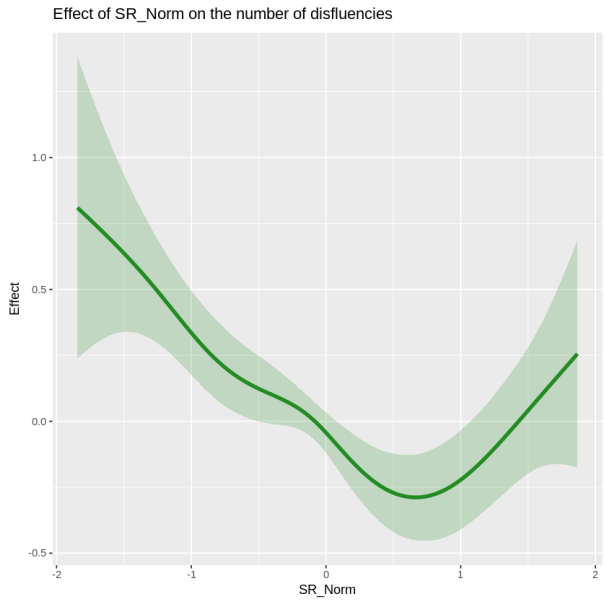

Figure 1: *The GAMM model of number of disfluencies and SR. The Y-axis represents the fitted effect, the X-axis represents the normalized SR*

The effect of the smooth SR_Norm is exemplified in Figure 1, which shows an overall negative and significant effect of SR on the number of disfluencies, as shown by the fitted falling curve. However, it should be noted that after a certain threshold, the effect stabilizes (see the region between the normalized SR 0.5 and 1). The fitted curve seems to show an apparent increase towards the higher SR, this may be due to the difficulty of maintaining speech at extreme SR levels. However, it should be noted that the confidence interval, represented in Figure 1 as the green band, crosses the zero, indicating statistical uncertainty on the estimation of the effect in this region. A possible cause for this effect may be the limited number of intonational units with such a high SR.

By extracting the model's parameters and using first derivatives, we calculated the minimum of the fitted curve, corresponding to the point where the negative trend changes. This corresponds to the normalized x-value $x = 0.66$. By applying in reverse this normalized minimum to the range of each speaker, we can estimate that this effect arises around 5.6 syl/s per speakers PS05, SP09, SP10 and 7.5 syl/s for speaker 1. To sum up, SR has a negative effect on the number of disfluencies until a certain high value is reached. From this value onwards, the effect is stable or difficult to establish.

3.2. Number of reductions

The second GAMM was fitted to predict the number of reductions as an depending on the normalized SR. The model's specification followed the same structure as the previous one and the assumptions were satisfactorily met for this case as well. The smooth for the SR had a significant effect, proving

Parametric	coefficients			
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.48	0.16	2.96	< .001
Approximate	significance	of smooths		
	edf	Ref.df	Chi.sq	p-value
s(SR_Norm)	2.37	2.98	28.07	< .001
s(Speaker)	2.79	3.00	44.70	< .001

Table 3: *GAMM model with number of reductions*

that it can predict reliably the number of reductions. In particular, by looking at Figure 2, we can observe that the effect of SR is positive throughout the model. However, it should be noted that also here after a certain threshold, we observe a change in the fitted curve. As SR_Norm increases towards zero, the effect gradually diminishes, stabilizing. This indicates that beyond a certain threshold of rapid speech, additional increases in SR yield only marginal increases in phonetic reduction. This threshold roughly corresponds to 4.8 syl/s for speakers 5, 9, 10 and 6.3 for speaker 1.

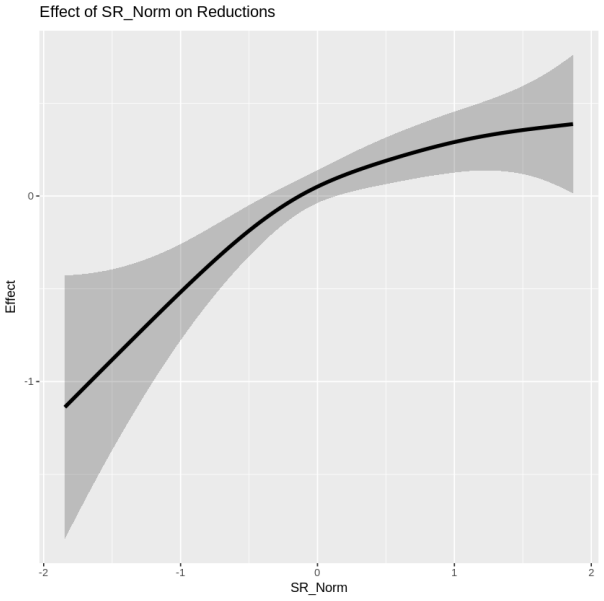

Figure 2: *The GAMM model of number of reductions and SR. The Y-axis represents the fitted effect, the X-axis represents the normalized SR*

4. Discussion

Our results indicate that variations in SR are associated with both the occurrence of disfluencies and phonetic reduction phenomena. These findings reveal complex and non-linear relationships that merit deeper theoretical consideration, since

they can reflect fundamental constraints in speech planning and execution.

Specifically, higher SRs tend to correlate with fewer disfluency occurrences (compare the discussion in [19] on whether SR can be seen as a precipitator of disfluencies in a clinical context). However, this effect appears to be limited to a certain threshold (on average about 6 syl/s), suggesting a saturation point. Beyond this saturation point, further increases in SR do not substantially reduce the number of disfluencies. The stabilization of this effect is particularly intriguing and may indicate a floor effect in disfluency production. Whether this effect is due to the scarcity of data in extremely high SR or it reflects a real cognitive constraint needs to be addressed in future and more robust studies.

We can be more confident on the effects of reduction. Higher SRs are generally linked to increased occurrences of phonetic reduction phenomena. These findings align with theoretical expectations in speech production models, whereby faster articulation rates create conditions that favor phonetic reduction processes, while slower, more careful speech tends to preserve fuller phonetic forms [20]. However, the rate of increase in reduction occurrences diminishes at very high SRs. This suggests that even when speakers increase their rate of speech, there is a limit to how much phonic material can be simplified. This ceiling effect may be due to a biomechanical limit to how much articulatory simplification can occur while maintaining intelligibility. With this in mind, the point around 4.8 and 6.3 syl/s (depending on the speaker) may represent an optimality point where speakers have maximized articulatory efficiency without compromising intelligibility.

5. Conclusions and Future Research

In this study we have examined the complex relationship between SR and disfluencies on the one hand, and SR and reduction on the other. Our results are in line with our general hypotheses, namely a higher SR correlates with a smaller number of disfluencies and a higher frequency of reductions; however, these results show a non-linear relationship, where the effects reach a threshold point. The presence of the latter might derive from cognitive factors. On a higher level, both results may reflect the interplay between cognitive planning and articulatory organization. Sustained speed may be made possible by a smaller cognitive load and a faster motor planning, which is reflected in a lower number of disfluencies. However, in the current state of knowledge, the highest SR values do not necessarily imply the complete disappearance of disfluencies, which are therefore unavoidable. Vice versa, slower processing and planning are detectable with a slower SR and with an increased number of disfluencies. To corroborate these points, we would need to include other factors in our models, such as syntactic and prosodic complexity, intonational units' length, accessibility of information and other cognitive factors (as advocated by [7, 19] among many others).

Besides these considerations, our study has highlighted the fact that these three dimensions are connected in a non-linear fashion, requiring further cognitive investigation to understand precisely what are the causes of the changes in trends.

Furthermore, as an initial investigation into the relationship between disfluencies, phonetic reductions and SR, this study does not allow us to establish the cause-and-effect relationship between these phenomena. Future research will further focus on the relationship between disfluency and reduction and see SR as a mediating variable between these two aspects of spontaneous speech production. A bigger corpus would also allow us to specify the current analyses for each disfluency or reduction type in order to ascertain whether these effects are cross-categorically consistent.

6. References

[1] R. J. Lickley, "Fluency and disfluency," *The handbook of speech production*, pp. 445–474, 2015.

[2] M. Ernestus and N. Warner, "An introduction to reduced pronunciation variants," *Journal of Phonetics*, vol. 39, no. SI, pp. 253–260, 2011.

[3] F. Chambers, "What do we mean by fluency?" *System*, vol. 25, no. 4, pp. 535–544, 1997.

[4] M. Ernestus, "Acoustic reduction and the roles of abstractions and exemplars in speech processing," *Lingua*, vol. 142, pp. 27–41, 2014.

[5] I. Finlayson, R. J. Lickley, and M. Corley, "The influence of articulation rate, and the disfluency of others, on one's own speech." in *DiSS-LPSS*, 2010, pp. 119–122.

[6] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, "Effects of disfluencies, predictability, and utterance position on word form variation in english conversation," *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1001–1024, 2003.

[7] L. Plug, "Phonetic reduction and informational redundancy in self-initiated self-repair in dutch," *Journal of Phonetics*, vol. 39, no. 3, pp. 289–297, 2011.

[8] M. Voghera, V. Mayrhofer, M. Ricci, F. Rosi, C. Sammarco *et al.*, "Il modokit. un identikit modale delle produzioni parlate e scritte dalle medie al biennio," in *Orale e scritto, verbale e non verbale: la multimodalità nell'ora di lezione*. Cesati, 2020, pp. 227–245.

[9] L. Schettino and F. Cutugno, "Quantifying and characterizing phonetic reduction in italian natural speech," *Languages*, vol. 10, no. 1, p. 14, 2025.

[10] P. Boersma, "Praat: doing phonetics by computer [computer program]," *http://www.praat.org/*, 2011.

[11] H. Sloetjes and P. Wittenburg, "Annotation by category-elan and iso dcr," in *6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2008.

[12] L. Schettino, "The role of disfluencies in italian discourse. modelling and speech synthesis applications," Ph.D. dissertation, Ph. D. dissertation, Università degli Studi di Salerno, 2022.

[13] J. Ginzburg, R. M. Fernández, and D. Schlangen, "Disfluencies as intra-utterance dialogue moves," *Semantics and Pragmatics*, vol. 7, no. 9, p. 64, 2014.

[14] M. Nespor and I. Vogel, *Prosodic phonology*. Walter de Gruyter, 1986, vol. 28.

[15] S. Shattuck-Hufnagel and A. E. Turk, "A prosody tutorial for investigators of auditory sentence processing," *Journal of Psycholinguistic Research*, vol. 25, pp. 193–247, 1996.

[16] D. R. Ladd, *Intonational phonology*. Cambridge University Press, 2008.

[17] N. P. Himmelmann, M. Sandler, J. Strunk, and V. Unterladstetter, "On the universality of intonational phrases: A cross-linguistic interrater study," *Phonology*, vol. 35, no. 2, pp. 207–245, 2018.

[18] S. N. Wood, *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC, 2017.

[19] J. Sawyer, H. Chon, and N. G. Ambrose, "Influences of rate, length, and complexity on speech disfluency in a single-speech sample in preschool children who stutter," *Journal of Fluency Disorders*, vol. 33, no. 3, pp. 220–240, 2008.

[20] B. Lindblom, "Explaining phonetic variation: A sketch of the h&h theory," in *Speech production and speech modelling*. Springer, 1990, pp. 403–439.