DOI: 10.17469/O2112AISV000009

EMANUELA GALLO, LOREDANA SCHETTINO

Acoustic characteristics of prolongations in spontaneous Italian speech[1]

This study concerns the analysis of segmental prolongations from an acoustic and functional perspective. Since the lengthening of sounds is a common feature of speech, recognizing disfluent-related occurrences is a complex operation and the main method commonly used is a "pragmatic approach" based on subjective perceptive judgment. However, there is a lack of investigations concerning acoustic characteristics that, in addition to duration, contribute to the perception of markedness and distinguish the portion of the segment identifiable as a prolongation from that identifiable in context as an unmarked part of the lexical element. From the results, it emerges that prolongation instances exhibit longer duration, reduced intensity, lower jitter and shimmer values, and higher Harmonics-to-Noise Ratio (HNR) measurements. Furthermore, their functional differences have been observed to correlate with different acoustic realizations.

Keywords: Prolongations, Disfluencies, Acoustic features, Functions.

1. *Introduction*

Human spontaneous speech is a complex activity that involves online planning, production, and reception processes. During these processes, speakers may produce phonetic cues to temporally suspend speech and gain extra time for planning or they may abandon utterances due to changes in plans and edit already uttered sequences considered somehow imprecise for the communicative intentions. These heterogeneous phenomena that interrupt the flow of speech making it discontinuous, are generally referred to with the term *disfluencies*. Early studies on these phenomena have been driven by the observation of their correlation with pathological conditions and aimed at distinguishing typical disfluencies from atypical disfluencies (a.o.; Johnson, 1961; Wingate, 1984, Zmarich, 2017). In fact, further studies have shown that disfluencies not only normally occur in typical speech but are not just occasional and individual production errors, but regularly occur in speech and are naturally involved in the economy of speech (Crocco, Savy, 2003), used as linguistic tools by speakers to control the quality of their production and to manage the online processes of speech planning, articulation, and reception (Allwood, Nivre & Ahlsén, 1990; Ginzburg, Fernández & Schlangen,

[1] This article is the result of the collaboration among the authors. However, for academic purposes only, Emanuela Gallo is responsible for sections 2, 3 and 5. Loredana Schettino is responsible for sections 1 and 4. Both are responsible for sections 6 and 7.

2014; Lickley, 2015; Voghera, 2017). Although different terminologies and classifications of disfluency phenomena have been developed according to specific research domains, interests and approaches, there is a certain agreement on Levelt's description of the general structure of disfluency events (Levelt, 1983) based on his proposed model of speech production (Levelt, 1989)[2]. According to this framework, speakers can monitor their own produced (overt) speech as well as their (inner) speech plan and thus detect and take action on issues either before or after articulation. Hence, two major subtypes of disfluencies have been distinguished, namely *repairs* and *hesitations*. Repairs or Backward-looking disfluencies, such as substitutions, insertions, and deletions, occur when speakers interrupt the ongoing speech and trace back to previously spoken material correcting and modifying it, or re-establishing continuity with the preceding utterance. Hesitations or Forward-looking disfluencies refer to a temporary suspension of flowing speech (Lickley 2015) and consist of a range of phonetic cues that allow speakers to gain extra time to deal with information selection, lexical retrieval, or other planning/production issues (Allwood et al., 1990; Ginzburg et al., 2014). Given that speech unfolds in real-time within diverse communicative contexts and conditions, hesitations are regarded as conversational tools that may be useful for speakers to gain extra time to manage their discourse and, at the same time, for listeners to process information. Their occurrence was observed to be mainly connected to moments that require a higher cognitive load in the speech planning process, such as the conceptualization and formulation of a new utterance, the search for a specific lexical item, and the selection of new information in discourse. These suspension devices may manifest through prospective repetitions, silent pauses, filled pauses (non-lexical and lexical fillers), and by prolonging speech sounds beyond their normal duration. As observed by Betz, Wagner (2016), lengthening is the first signal of hesitation, the primary measure a speaker can apply to solve problems in speech planning. The term *prolongations* has been used in the literature to refer to disfluent lengthenings which have been described as "phones that are longer than would be expected in normal, fluent speech" (Eklund, 2004: 163) or "a marked prolongation of one or more phones, resulting in a longer than average duration of syllables and words [...] This coincides with a local reduction in speaking speed which is not expected by the listener, causing an impression of disfluency and hesitation" (Betz, 2020: 14). Considering that the perception of the length of segments as being *marked* depends on the context of occurrence rather than actual duration, the identification of prolongations is a non-trivial operation[3]. This study aims to investigate to what

[2] Various models and theories of speech production have been proposed to account for the well-coordinated speech movements underlying the production of spoken language. However, it has been acknowledged that theories still struggle to properly account for all the properties of these movements. A collection of studies exploring the dynamics of speech motor control can be found in the volume "Models and theories of speech production" edited by Gafos, van Lieshout (2020).

[3] Grounded on the assumption that in spoken communication, speakers tend to make use of different available modes (vocal as well as gestural) in a cooperative and integrated manner, it has been argued that speech phenomena should be analyzed with a multimodal perspective to get a clearer picture of

extent acoustic characteristics, beyond duration, contribute to the perception of markedness of disfluent lengthenings in speech as well as to examine if different functions of prolonged speech segments influence acoustic parameters changes.

The article is structured as follows: § 2 presents the literature concerning prolongations and filled pauses; § 3 presents the general and specific aims of the study; § 4 describes the linguistic data, the annotation process, and the parameters of the linguistic analysis; in § 5 and § 6 the results of the analysis are reported and discussed.

2. *Related work*

2.1 Prolongations and filled pauses

Segmental prolongations have been largely described as a kind of stuttering behavior which, opposite to repeated movements, do not consist of atypical movements but of absence of typical movements (Onslow, 2023). However, this type of disfluency phenomena is also one of the most common form of non-pathological speech disfluency (Eklund, 2001, 2004; Williams, 2022) and frequently co-occur with other types of disfluent phenomena in everyday speech, reflecting the transitional speakers' lack of motor control during speech planning, i.e., when speakers need time to overcome some execution difficulty, or serving as a strategy to gain additional time to solve the problem in planning processes (Gósy, Eklund, 2018). Since the lengthening of sounds is a common feature of speech serving various functions, i.e., phrase-finality, accentuation, back-channeling and hesitation (Betz, 2020), a key feature to distinguish prolongations from other types of lengthening lies in their pitch contour and occurrence within syllables: disfluent lengthenings are generally realized with lower pitch range and flat pitch contour and occur in long vowel nuclei as well as in sonorant codas (Betz, 2020). Nevertheless, prolongations are hard to classify in a rule-based way as syllables vary in duration in fluent speech according to context making unclear how long they must be in order to be perceived as a hesitant disfluency element. According to Lickley (2015), a possible solution consists of adopting a "pragmatic approach", based on subjective perceptive judgment, supported by reliability measures of inter-annotator agreement. Even if duration appears to be the main acoustic cue to be used in prolongation recognition, it is not always the decisive factor. Firstly, it is important to consider that perceived prolongation is dependent on speech rate and thus needs to be calibrated against a durational framework, rather than being defined in absolute durational terms. Secondly, there is significant variation among listeners in their temporal thresholds for identifying prolongations: listeners may perceive elongated

their role and production. Different approaches have been proposed to tackle this, e.g., some studies rely on multimodal transcription systems (Kosmala, 2024), whereas other studies carry out the analyses at the speech and gestural levels independently and then merge these levels to observe their correlation in the speech flow, as in Schettino, Campisi & Origlia (2024).

sounds along a continuum wherein prolonged segments of a word are intentionally produced and hence not regarded as instances of disfluency (Gósy, Eklund, 2018). Therefore, several other factors may contribute to the perception of markedness and influence prolongation recognition, such as articulation rate, fundamental frequency, intensity, phonetic context, segment quality, and position in the phrase (Gósy, Eklund, 2018). Most of these markedness-contributing factors have been systematically analyzed for filled pauses, which similarly to prolongations, also create a suspension and delay production through vocalization and duration, distinguishing themselves from other hesitation phenomena (Eklund, 2004). Regarding their manifestations, filled pauses contain non-lexical material and are typically comprised of a central vowel, optionally followed by a nasal, e.g., *eeh, ehm, mhh* (Di Napoli, 2020). Although there is evidence indicating that filled pauses share standard acoustical features across different languages, such as a flat pitch and stable formant frequencies (Hamzah, Jamil & Seman, 2012), studies have shown that there is variability in the way these phenomena are produced, influenced by both language-specific factors and individual speaker characteristics (Giannini, 2003; Jabeen, Wagner, 2023). Studies on production mechanisms of filled pauses were conducted within the context of Japanese spontaneous speech, both in monologue and conversation settings, concentrating on the acoustic differences between vowel sounds in vocalizations and those belonging to lexical items. Maekawa, Mori (2017) highlighted that the voice quality of filled pauses seems to be used by speakers to transmit various paralinguistic information and that there is a systematic difference in voice quality between vowels in filled pauses and those in lexical items which correlates to different functions in speech. They showed that vowel sounds in filled pauses had relatively "softer" phonation, as indicated by higher spectral tilt and H1-H2 ratio values, i.e., the difference between the first and second harmonic, and were also unstable, as indicated by larger jitter values, i.e., the cyclic variation of the fundamental frequency of a vocal sound. Additionally, they were also marked by lower intensity, lower F0, and higher F1.

Similarly, Li, Ishi, Fu & Hayashi (2022) investigated the differences in filled pauses between Japanese native speakers and Chinese learners of Japanese as L2 employing both prosodic and voice quality measurements. Their findings indicate that vowels of filled pauses produced by Japanese speakers exhibit longer duration, lower intensity, and more aperiodicity, whereas Chinese learners of L2 Japanese show longer duration, lower F0 mean, and a tenser phonation type to distinguish filled pauses from ordinary lexical items. Moreover, their filled pause production patterns are influenced by their native language as indicated by features such as intensity, H1-A3 ratio, i.e., the difference in frequency between the first harmonic and the third formant, jitter and shimmer, i.e., the cycle-to-cycle variation of the vocal fold in relation to the amplitude of the sound wave.

These investigations have demonstrated that vowel differences in filled pauses from other lexical items involve both prosodic and vocal quality parameters, with duration and intensity contributing significantly to the distinction while parameters

linked to vocal quality such as jitter, shimmer, HNR (which measures the relative amount of additive noise in the vocal signal and the degree of acoustic periodicity), and H1-H2 ratio played a secondary role. Given these observations, it is reasonable to hypothesize that voice quality features could contribute to the perception of markedness of prolongation phenomena.

2.2 Cross-linguistic studies on prolongations

The literature on the characterization of segmental prolongations has mainly concerned their position within the word being prolonged, the relevant part of speech, segmental content, duration and pitch (see Tab. 1).

Studies concerning prolongations in different languages show both common patterns and language-specific traits. Regarding the distribution within the initial/medial/final segments of words, prolongations appear to be strongly dependent on syllable structure and language-specific phonotactic rules. Indeed, in the series of studies conducted by Eklund on prolongations in different languages, he observed differences and similarities in the word positioning of prolongations in relation to morphological complexity as well as phonotactic rules and syllable structure complexity. First, he proposed the "morphology matters" hypothesis (Eklund, 2004), which postulates that the distribution of prolongations is related to the morphological complexity of the language under consideration. This hypothesis was supported in Swedish and American English (Eklund, Shriberg, 1998), two languages with similar morphological complexity and which show an identical distribution of prolongation in initial, medial, and final positions, contrasting with Tok Pisin pidgin, which is considered much less complex and show a prevalence of initial and final occurrences. On the one hand, subsequent works on Japanese, Mandarin, and Hungarian support this hypothesis (Den, 2003; Lee, 2004; Gósy, 2017). On the other hand, investigations on German (Betz, 2017) presented an exception, as the distribution of prolongations resembled that of Tok Pisin, Japanese, and Mandarin despite showing morphology traits closer to Swedish, American English and Hungarian. Hence, Eklund proposed the "phonotactics matters hypothesis" (Gósy, Eklund, 2018). This new framework was further supported by a study on Hebrew (Silber-Varod, 2019), as well as Italian speech data, where prolongations occur almost exclusively in the final position of the word (Schettino, 2023). The conclusions that were drawn from the above-mentioned studies are that, although morphology and phonotactics may play some role in what position of the word prolongations might appear, other factors must also play a role.

For what concerns the part of speech of words, prolongations tend to occur more frequently on function words rather than content words in Swedish, Tok Pisin, German and Hungarian, while no clear distinction was found between Mandarin, Japanese, and Italian (Di Napoli, 2020; Schettino, 2023).

Regarding segmental content, all types of segments might be subject to prolongation, but vowels and sonorants are generally more prone, within language-specific phonological constraints. In German (Betz, 2017), sonorants are the target

of prolongations, followed by the combination of diphthongs and long vowels. In Hungarian, Hebrew, and Turkish (Altiparmak, Kuruoğlu, 2018), prolongations involve more vowels than consonants. The same happens in Mandarin (Lee, 2004), where vowels and vowel-coda pairs are more frequently prolonged than consonants. Likewise, in Italian (Schettino, Eklund, 2023), most cases of prolongations are made by the lengthening of vowel sounds, occurrences of lengthening of word-final consonants followed by a schwa sound, diphthongized and lengthened final vowels with schwa sound, as well as lengthening of continuous consonants or nasals, and rare cases of word-final vowels plus lengthening of the nasal sound. Maybe these cases of schwa insertion in word-final prolongations led to the perception and classification of prolongations as a particular type of filled pause in Italian studies (see Schettino, Eklund, 2023).

In terms of duration, prolongations are generally longer on vowels than on consonants in both Hungarian and Italian whereas no significant durational difference between vowels and consonants was found in Hebrew. For what concerns the other languages under investigation, prolongations have been analyzed in comparison to filled pauses, which were found to have greater average duration than lengthenings in different languages, including European Portuguese (Moniz, Mata & Viana, 2007), Swedish, Tok Pisin, German, and Italian.

Nevertheless, systematic investigations concerning acoustic characteristics that, in addition to duration, contribute to distinguishing the portion of the segment identifiable as a marked prolongation is still lacking.

Table 1 - *Summary of studies on prolongations and filled pauses with analyzed parameters*

Hesitation Type	Language	Studies	Position within the word	Part of speech word	Segmental Content	Duration	F0, Pitch	Intensity	Spectral Tilt	Jitter and Shimmer	HNR	F1,F2
PRL	Swedish, Tok Pisin	Eklund, R. (2001).	✓	✓		✓						
PRL	English	Shriberg, E. E. (1994).	✓									
PRL	Japanese	Den, Y. (2003).	✓	✓								
PRL	Mandarin	Lee, T.-L., He, Y.-F., Huang, Y.-J., Tseng, S.-C., & Eklund, R. (2004).	✓	✓								
PRL	Hungarian	Gosy, M., & Eklund, R. (2017).	✓	✓	✓	✓						
PRL	Hebrew	Silber-Varod, V., & Gosy, M., & Eklund, R. (2019).	✓		✓	✓						
PRL	German	Betz, S., Eklund, R., & Wagner, P. (2017).	✓	✓	✓	✓	✓					
PRL	Italian	Schettino, L., & Eklund, R. (2023).	✓	✓	✓	✓						
PRL, FP	Italian	Di Napoli, J. (2020).				✓						
FP	Japanese, Chinese	Li, X., Ishi C.T., Fu, C. & Hayashi, R. (2022).					✓	✓	✓	✓	✓	
FP	Japanese	Maekawa, K., & Mori, H. (2017).								✓		✓
FP	Punjabi	Jabeen, F., & Wagner, P. (2023).					✓					✓

3. *Research aim*

Besides the main acoustic cue of duration, other factors may contribute to perceiving prolonged sounds as disfluent-related phenomena. Given this framework, the present study aims to characterize the distinction between the disfluent portion of a prolonged segment (e.g. th**e<ee>**) from that identifiable in context as an unmarked part of the lexical item (e.g. th**e**<ee>) by deepening the analysis of phonetic-prosodic parameters as well as parameters related to voice quality, further supporting the recognition of the linguistic value of prolongations serving as a tool for managing spoken productions and the reliability of the functional approach commonly adopted for their identification.

Hence, the proposed study aims to address the following questions:

1. To what extent do acoustic characteristics, beyond duration, contribute to the perception of markedness of disfluent lengthenings?
2. How do different functions of prolonged speech segments influence acoustic parameter changes?

4. *Methodology*

4.1 Corpus and dataset

The analysis was conducted on the "Modokit-FROG" corpus (Sarro, 2023) consisting of audio-visual recordings of 10 Italian speakers (5 females, 5 males; Mean age = 25.1 SD = 1.91). As the corpus was specifically collected to identify and analyze Manner expressions through speech and gestures, speakers are not reported (nor observed) to be affected by speech pathologies. Participants come from different regions, namely Abruzzo (5), Lazio (2), Campania (2), and Apulia (1), and are all students at the University of L'Aquila. Following the Modokit protocol (Voghera, Mayrhofer, Ricci, Rosi, & Sammarco, 2020), the data collection involved eliciting both written and oral productions using the "frog stories" method, which consists of input taken from the picture book "Frog, ¿where are you?" by Mercer Mayer (1969). Each recording session comprised a speaker who actively participated by telling the story to a silent interlocutor, who was unaware of the story's content. The initial group, consisting of four males and one female, described the twenty-nine images first orally and then provided written descriptions (PS task). The second group, comprising four females and one male, did the opposite, initially writing the frog story and then narrating it orally to the addressee (SP task).

4.2 Annotation

Firstly, disfluency phenomena have been annotated on the ELAN software (Sloetjes, Wittenburg, 2008) employing a multilevel annotation scheme that provides information at formal and functional levels. The first level concerns the macro-structures of the disfluencies in which several regions are delineated following Shriberg's (1994) disfluency model, namely, the region to be repaired (RP), the

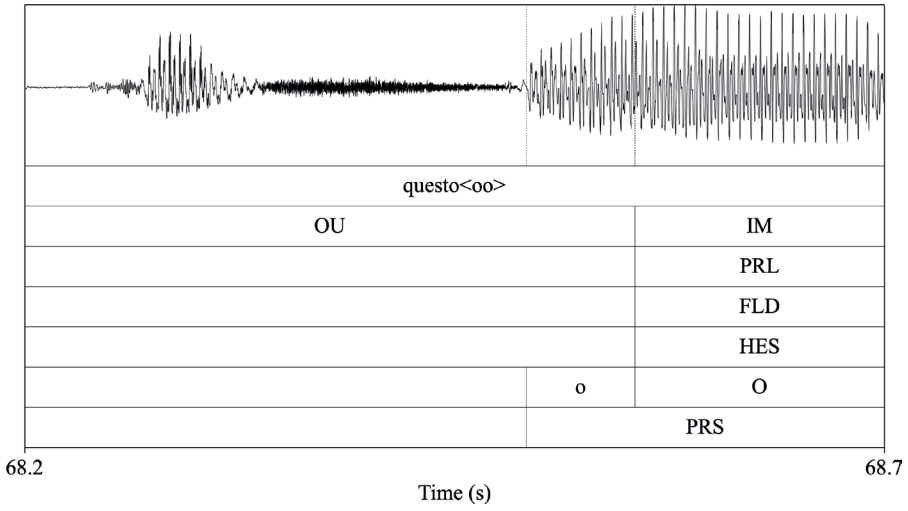
repaired one (RS), and the one in which the delay occurs (IM). The second level focuses on the micro-structure of disfluency, providing detailed information about the specific type of disfluency phenomenon. Among other phenomena, disfluency-related prolongation occurrences (PRL) were identified as marked lengthening of segmental material (Eklund, 2004; Betz, 2020) without establishing a minimum duration threshold and relying only on auditory judgment, i.e., once identified the unmarked-sounding portion of the lengthened word, the remaining portion was labelled as prolonged segment. At the third level, each item is assigned its macro-function, categorized as either Backward-Looking (BLD) or Forward-Looking disfluency (FLD) (Ginzburg et al., 2014). 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 in retrieving the target word.
- *Structuring* (STR) for items involved in structuring the discourse on a syntactic and informative level.
- *Focusing* (FOC) for items preceding semantically relevant elements.
- *Hesitative* (HES) is the label assigned to items that primarily serve the basic planning function of speech and it is applied to instances where no other overlapping function can be identified[4].

The annotations were performed by two expert judges and were evaluated by measuring the values of Cohen's k, i.e., 0.82 for the type of disfluency phenomenon and 0.77 for specific functions, standing respectively for "high agreement" and "substantial agreement" (Landis, Koch, 1977). Subsequently, all instances of prolongation were further labeled on Praat (Boersma, Weenink, 2021) at two separate levels (see Fig. 1): on one tier, the entire segment subject to prolongation was labeled as Prolonged Segments (PRS). Then, each PRS was further divided into a prolongation segment, which refers to the marked extension of segmental material corresponding to PRL, and a segment identified as unmarked (word segment, w).

[4] The scheme is described in detail in Schettino (2022).

Figure 1 - *Praat annotation system*



4.3 Linguistic and statistical analysis

At this stage, the analysis focused on prolongations of vocalic segments. For these, each PRL and w segments were analyzed concerning the following acoustic parameters:

- Duration (ms)
- Intensity (midpoint, z-scored)
- Fundamental frequency (midpoint, Normalized 100Hz)
- Formant frequencies (midpoint, Lobanov normalized, "PhonR")
- Jitter (local)
- Shimmer (local)
- Harmonics-to-Noise Ratio (mean, dB)

All the measurements selected as the object of investigation and the information encoded following the annotation scheme were extracted using a Praat script. The acoustic parameters and prolongation functions were processed and subjected to statistical analysis using the R software (R Core Team, 2021). To control for individual variability, the values obtained for each parameter were standardized by computing the z-score. For the formant values, the z-score method was employed, also referred to as the Lobanov normalization (Lobanov, 1971).

For statistical analysis, Linear Mixed Models ('lme4' package, Bates, Maechler, Bolker & Walker, 2015) were fitted with Speaker as random effect. Different models were built for each acoustic parameter selected as the dependent variable. Independent variables were: Prolonged Segment (levels: PRL, WS) and sociolinguistic factors, i.e., sex, age, and place of origin. For the PRL subset, the Function was introduced as a fixed factor. A post-hoc analysis was also conducted (emmeans package, Lenth,

Ingmann, Love, Buerkner & Hierve, 2018). The results yielded by the statistical analysis were interpreted taking into account descriptive measurements extracted with the "summarySE" function (Rmisc package, Ryan M. Hope, 2012).

5. *Results*

5.1 General results

In the selected dataset, consisting of 91 minutes of recordings, 641 phenomena of prolongation of word-final segments were identified. As previously mentioned, the study focused on prolongations of vocalic segments, which were analyzed with reference to standard Italian vocalic sounds, resulting in a total of 549 occurrences. Table 2 outlines the number of prolongation instances per speaker. On average, there were about 5 phenomena of prolongation per minute across all speakers, with a mean of 5,91 occurrences in the PS task and a mean of 5,87 in the SP task. Given the high variability among speakers in terms of the occurrence of prolongations, mixed models with the speaker as a random effect were deemed appropriate to support the analysis.

Table 2 - *Duration, number of words, average words per minute, prolongation occurrences, and average prolongation occurrences per minute*

n	*task*	*Dur(m)*	*word*	*word/m*	*PRL(n)*	*PRL/m*
1	PS	10,42	800	76,7	39	3,74
2	PS	10,67	1034	96,9	40	3,75
3	PS	12,78	1717	134,3	103	8,06
4	PS	15,15	1315	86,7	113	7,46
5	PS	10,07	816	81	66	6,56
6	SP	8,92	1270	142,3	9	1,01
7	SP	5,83	656	112,5	49	8,40
8	SP	3,78	385	101,8	5	1,32
9	SP	6,98	790	113,1	66	9,45
10	SP	6,45	685	106,2	59	9,15

5.2 Phonetic-prosodic parameters

In Fig. 2, the results for duration are depicted. As expected, prolongation segments are significantly longer than the unmarked portion of the prolonged segments as they consist of the lengthening of segmental material (see Tab. 3 and Tab. 4).

Figure 2 - *Duration values in word (w) and prolongation segments (PRL)*

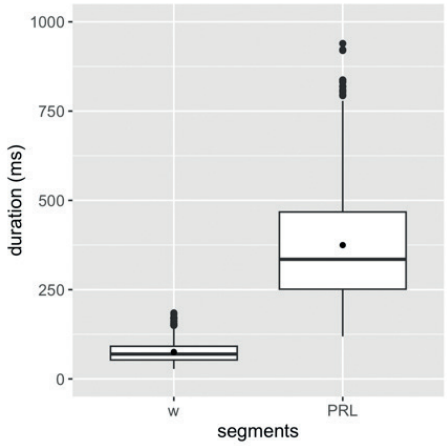


Table 3 - *Mean, standard deviation, standard error of the mean, and confidence interval of duration values in w and PRL*

Segm.	*Duration (ms)*	*sd*	*se*	*ci*
w	75	29	1.26	2.48
PRL	376	167	7.1	14.07

Table 4 - *Pairwise comparison among the levels of the Prolonged Segment for the duration model*

Contrast	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	-301	7.13	-42.200	<.00

Regarding intensity, Fig. 3 shows that there is a slight decrease in prolonged segments (see Tab. 5 and Tab. 6).

Figure 3 - *Z-scored intensity*

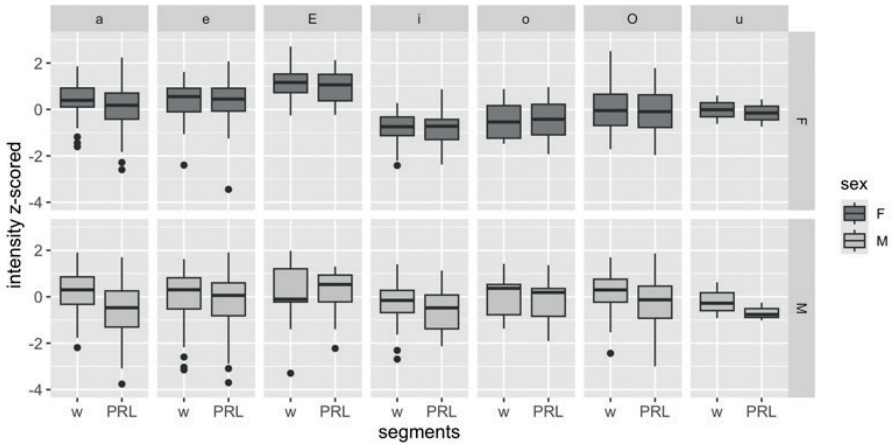


Table 5 - *Mean, standard deviation, standard error of the mean, and confidence interval of intensity values in w and PRL*

sex	*Segm.*	*Intensity midpoint (dB)*	*sd*	*se*	*ci*
F	w	61.24	5.25	0.33	0.66
F	PRL	60.73	5.23	0.33	0.65
M	w	60.74	5.04	0.28	0.56
M	PRL	58.49	5.43	0.31	0.61

Table 6 - *Pairwise comparison among the levels of the Prolonged Segment for the intensity model*

Fixed effects	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	-0.0236	0.008	-2.639	0.02

As shown in Fig. 4 and Tab. 7, a slight, though not statistically significant, decrease is also observed in fundamental frequency values of prolonged segments.

Figure 4 - *Fundamental frequency in word and prolongation segments*

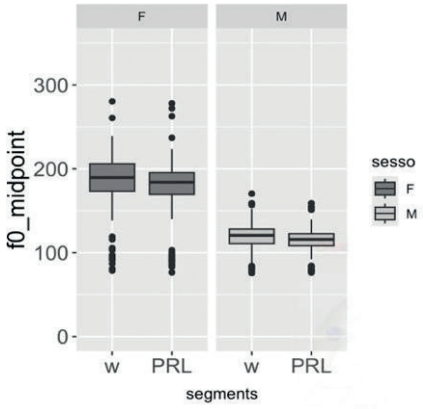


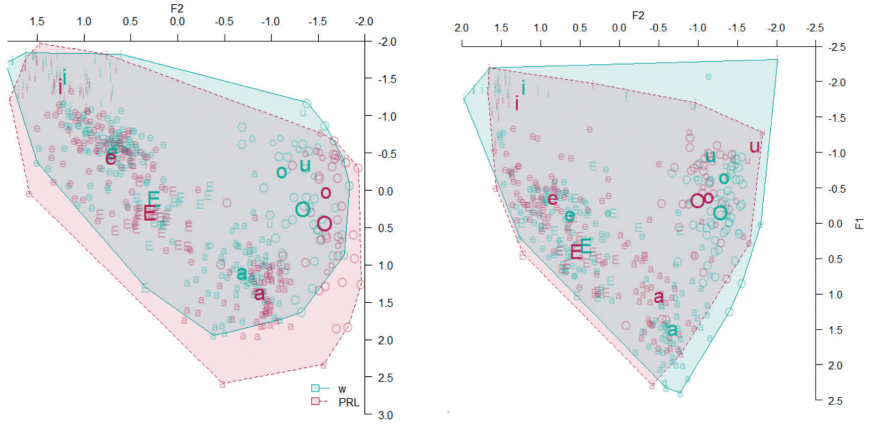
Table 7 - *Mean, standard deviation, standard error of the mean, and confidence interval of F0 midpoint values in w and PRL*

sex	*Segm.*	*F0 midpoint (Hz)*	*sd*	*se*	*ci*
F	w	189.20	41.89	3.02	5.96
F	PRL	187.9	32.95	2.18	4.30
M	w	122.82	32.24	2.45	4.83
M	PRL	115.51	13.22	0.95	1.87

5.3 Voice quality parameters

Concerning formant frequencies, no significant differences are found in the data between portions of prolongations and lexical items. As shown in Fig. 5, there is an overlap between the realizations of vowel elements belonging to the unmarked lexical item preceding the segment identified as prolongation in both men and women.

Figure 5 - *Formant frequencies of all vowels for men (left) and women (right)*



To examine the periodic disturbance of vocal fold vibration, jitter and shimmer measurements were utilized. As illustrated in Fig. 6 (see also Tab. 8 and 9), results indicate a decrease in jitter during prolongations when compared to previous lexical items in both male and female speakers for all vowel segments. The same decrease is shown in Fig. 7 (see also Tab. 10 and 11) for shimmer values.

Figure 6 - *Z-scored Jitter Local*

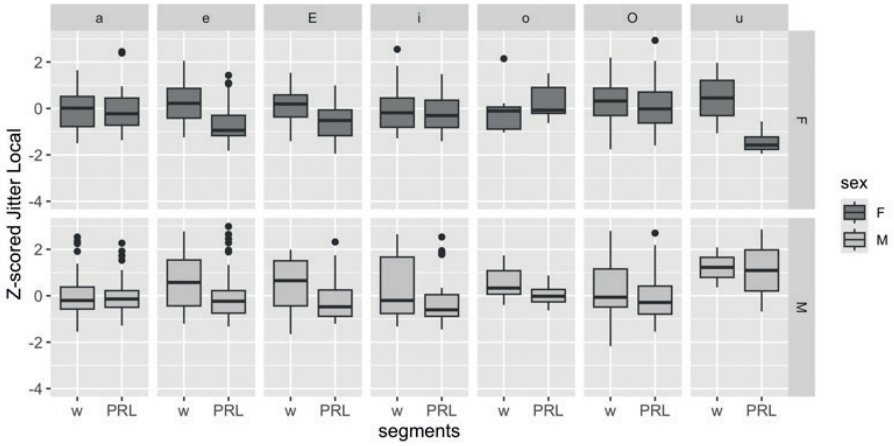


Table 8 - *Mean, standard deviation, standard error of the mean, and confidence interval of jitter local values in WS and PRL*

sex	*Segm.*	*Jitter local*	*sd*	*se*	*ci*
F	w	0.013	0.011	0.0007	0.0015
F	PRL	0.009	0.012	0.0007	0.0015
M	w	0.020	0.020	0.0016	0.0032
M	PRL	0.013	0.013	0.0011	0.0023

Table 9 - *Pairwise comparison among the levels of the Prolonged Segment for the jitter local model*

Fixed effects	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	0.269	0.08	3.104	<.00

Figure 7 - *Z-scored Shimmer Local*

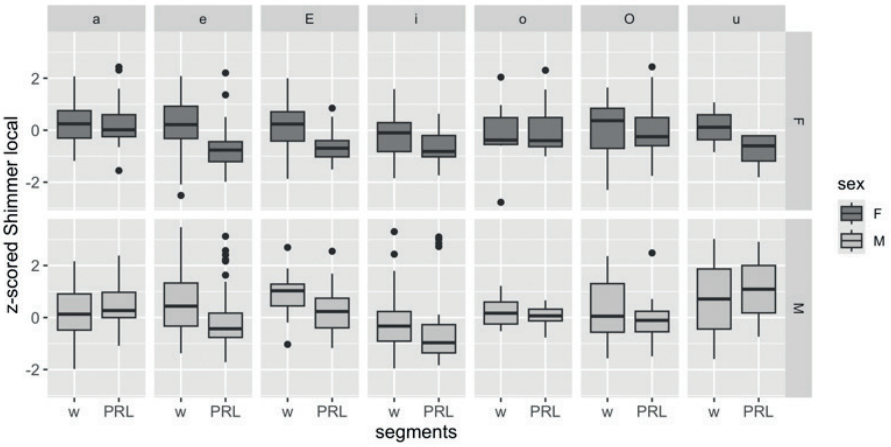


Table 10 - *Mean, standard deviation, standard error of the mean, and confidence interval of shimmer local values in w and PRL*

sex	*Segm.*	*Shimmer local*	*sd*	*se*	*ci*
F	w	0.068	0.04	0.003	0.006
F	PRL	0.045	0.03	0.002	0.004
M	w	0.081	0.06	0.004	0.009
M	PRL	0.063	0.05	0.003	0.007

Table 11 - *Pairwise comparison among the levels of the Prolonged Segment for the shimmer local model*

Fixed effects	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	0.327	0.108	3.26	<.01

As reported in Fig. 8, results indicate a statistically significant difference in HNR measurements between prolongations and word segments. The mean HNR is higher in prolongations than in word segments (see Tab. 12 and Tab. 13).

Figure 8 - *Mean HNR values*

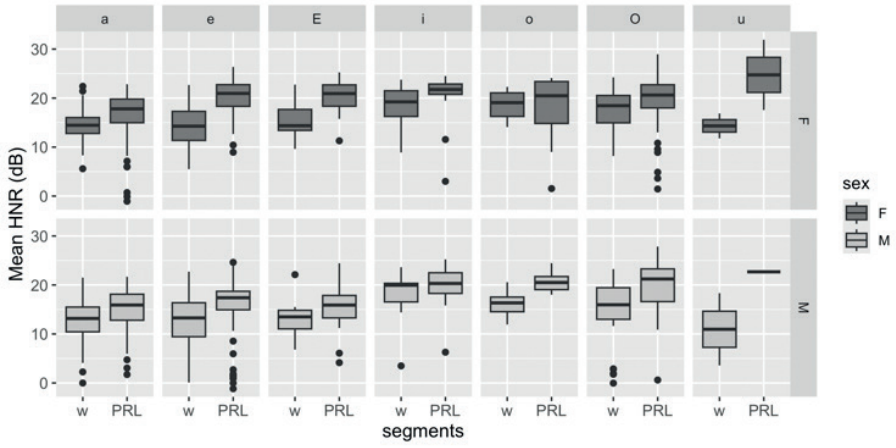


Table 12 - *Mean, standard deviation, standard error of the mean, and confidence interval of HNR values in w and PRL*

sex	*Segm.*	*Mean HNR*	*sd*	*se*	*ci*
F	w	15.48	4.24	0.29	0.58
F	PRL	19.15	5.22	0.33	0.66
M	w	13.19	5.55	0.39	0.78
M	PRL	16.34	5.81	0.40	0.79

Table 13 - *Pairwise comparison among the levels of the Prolonged Segment for the HNR model*

Fixed effects	*Estimate*	*SE*	*t-ratio*	*p. value*
w-PRL	-0.179	0.027	-6.444	<.00

5.4 Prolongation functions

All instances of prolongation of word-final vowels were analyzed in terms of functions based on the context of their occurrence. As shown in Fig. 9, these phenomena are primarily associated with the basic planning function (HES: 63 %), to a lesser extent connected with the discourse structuring function (STR: 22 %), lexical search (WS: 9 %), and the marking of semantically relevant elements (FOC: 6 %).

Figure 9 - *Prolongation functions frequency*

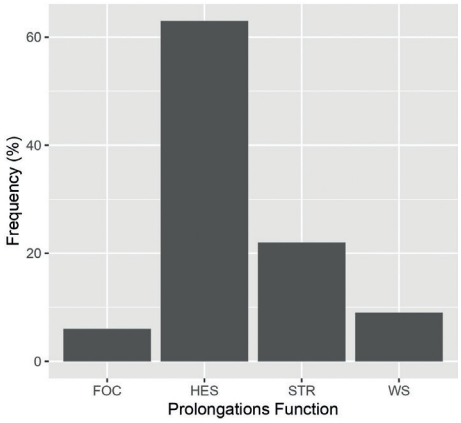


Figure 10 - *Phonetic-prosodic parameters of prolongations per function*

Figure a - *Duration*

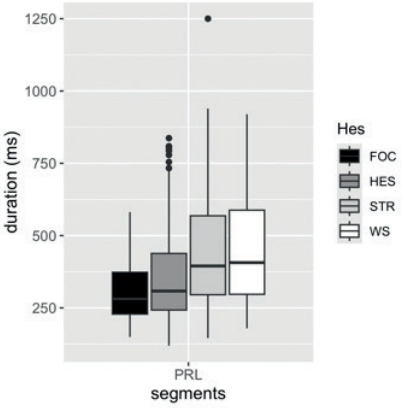


Figure b - *Intensity*

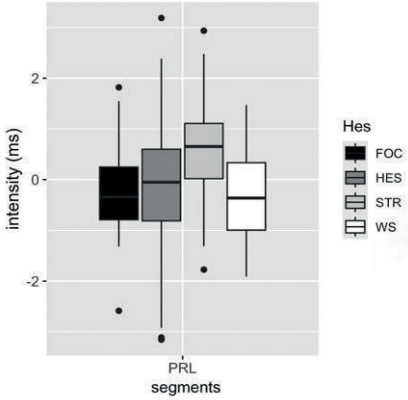
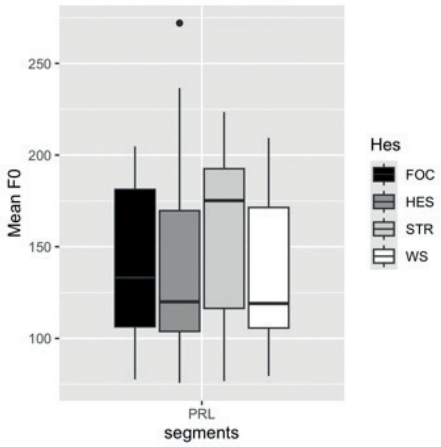


Figure c - *Mean Fundamental Frequency*



The analysis showed that the duration of prolongations varied significantly depending on the function they served within the speech. Figure 10a shows that prolongations exhibited a significantly longer duration when associated with the Word Searching (WS) function (Est = 136.23, SE = 36.02, t value = 3.782, p = 0.0002) and prolongations linked with the Structuring (STR) function (Est = 129.45, SE = 31.38, t value = 4.126, p <.00) as compared to those associated with Hesitative. Interestingly, prolongations associated with Focusing (FOC) function were even shorter (see Tab. 14).

Table 14 - *Mean, standard deviation, standard error of the mean, and confidence interval (default 95 %) of duration values per prolongations function*

Function	*Duration*	*sd*	*se*	*ci*
FOC	308	112	19.5	39.7
HES	350	147	7.94	15.6
STR	438	205	18.6	36.9
WS	447	169	23.7	47.6

As illustrated in Fig. 10b, prolongations linked with the structuring function showed significantly higher intensity values (Est =0.961, SE = 0.15, t value = 6.375, p <.00). Prolongations associated with hesitative function had slightly lower intensity values, closely followed by those associated with focusing function. Finally, prolongations associated with the word searching function had the lowest intensity values (see Tab. 15).

Table 15 - *Mean, standard deviation, standard error of the mean, and confidence interval (default 95 %) of intensity midpoint values per prolongations function*

Function	*Intensity*	*sd*	*se*	*ci*
FOC	58.2	4.9	0.8	0.8
HES	58.5	5.3	0.2	0.5
STR	63	4.4	0.4	0.8
WS	57	4.8	0.6	1.3

For what concerns fundamental frequency, as illustrated in Fig. 10c, prolongations linked with the Structuring function exhibited the highest mean F0 values (Est = 3.198, SE 0.55, t value = 5.761, p <.00), whereas prolongations associated with generic planning showed significantly lower mean F0 value (Est = 1.8541, SE = 0.5265, t value = 3.521, p = 0.00). Prolongations involved in lexical retrieval had the lowest mean F0 value (Est = 1.6292, SE = 0.6471, t value = 2.518, p = 0.01) (see Tab. 16).

Table 16 - *Mean, standard deviation, standard error of the mean, and confidence interval (default 95 %) of mean F0 values per prolongations function*

Function	*Mean F0*	*sd*	*se*	*ci*
FOC	161.1	33	4.7	9.6
HES	150.3	48	2.6	5.2
STR	177.5	51	4.8	9.6
WS	146.1	34	4.2	8.4

6. *Discussion*

Based on the findings presented in this study, statistically significant differences concerning phonetic-prosodic and voice quality parameters were observed between the vocalic segments identified as unmarked and the markedly prolonged segments. Firstly, the increase in duration observed in prolonged segments aligns with expectations, as duration serves as a primary acoustic cue for identifying prolongation instances.

As speakers elongate sounds, there is typically a decrease in the muscular effort required for articulation, leading to an energy reduction and lower intensity values during prolongation production. Moreover, lower jitter and shimmer values as well as a higher HNR ratio indicate a lower presence of noise and a higher level of periodicity in the speech signal during prolongation compared to vowels in regular lexical items.

These findings suggest that, on average, speakers tend to maintain a more periodic and stable voice during prolongation production possibly due to articulatory economy, which refers to the tendency of speakers to produce speech sounds with minimal effort during moments of uncertainty (Stepanova, 2007).

Also, the lack of significant statistical difference between word-final unmarked segments and prolongation segments in fundamental frequency and formant frequencies could suggest that speakers maintain similar vocal settings while lengthening. Furthermore, results on fundamental frequency align with prior research on hesitation markers in Italian discourse (Schettino, Betz, Cutugno, & Wagner, 2021a), in which prolongations typically displayed a flat pitch contour and had a narrower pitch range compared to other hesitation phenomena. However, the absence of variation in F0 may also be attributed to the relatively small portion identified as unmarked lexical items compared to prolongation segments.

Regarding formant frequencies, the data do not reveal any tendency to centralization, i.e., a phonetic phenomenon where the articulation of a vowel sound shifts towards a more centralized position in the oral cavity. Generally, articulatory economy and efforts reduction would be consistent with centralization and thus, changes in formant values. However, the data are consistent with previous studies on Italian speech, which noted only occasional instances of prolongations realized by the elongation of word-final vowels with a schwa sound and that timbric realization of prolongations can be traced back to articulatory models of the phonological inventory rather than to articulatory economy (Schettino, Eklund, 2023; Giannini, 2003; Schettino, Betz, Cutugno, Wagner, 2021b).

Moreover, to further understand the role of prolongation in the management of speech, the analysis concerned the variation of phonetic-prosodic parameters according to prolongations specific functions, namely, Word Searching, Structuring, Focusing, and Hesitative.

Prolongations associated with the word searching (WS) function tended to be longer as they suspend speech delivery for more time, which leads to acknowledging the search for and the selection of a specific lexical item as a more time-consuming process. In addition, these prolongations showed the lowest intensity and F0 values, aligning with prior research on hesitation markers in Italian discourse (Schettino et al., 2021a), in which lower pitch was found to characterize prolongations as well as other hesitations associated with Word Searching function. Prolongations associated with structuring (STR) function, involved in the organization of information, were characterized by higher duration as long as higher intensity and fundamental frequency. Higher F0 values also characterize prolongations with focusing (FOC) function, which, however, were also characterized by shorter mean duration and, on average, lower intensity values. Finally, when prolongations co-occurred with the Hesitative (HES) function, they exhibited lower duration, intensity as well as F0 values.

7. *Conclusions*

To conclude, the findings presented in this study suggest that the production of markedly prolonged segments corresponds to a reduction in articulatory effort,

which is consistent with lower intensity values, and a more periodic and stable voice characterized by lower jitter and shimmer values and higher HNR measurements.

Furthermore, the phonetic-prosodic features of these phenomena have been observed to correlate with functional differences. Results show that more time is used either for the retrieval of specific lexical items or for marking information structure. However, F0 values are found to distinguish items involved in word searching or generic planning (Word Searching and Hesitative) from those involved in discourse structuring (Structuring and Focusing), respectively lower and higher values. This supports the idea that prolongations play a role in marking discourse, with their acoustic realization reflecting different specific functions.

The findings reported in this work contribute, to a certain extent, to add pieces to the general picture in understanding the perception of markedness of disfluency-related lengthenings, making them systematically recognizable from other types of lengthening in speech. Further investigations may concern other factors that influence the perception of disfluent prolongations, such as speakers' speech rate, prolongations' position in the phrase, and the surrounding context, taking into consideration when these phenomena co-occur with other disfluency markers, like silent pauses, filled pauses, and repetitions. Also, the characterization of these phenomena may be supported by further investigations from a multimodal perspective.

References

Allwood, J., Nivre, J. & Ahse E. (1990). Speech management – on the on-written life of speech. In *Nordic Journal of Linguistics*, 13, 3-48. https://doi.org/10.1017/S0332586500002092

Altiparmak, A., Kuruoğlu, G. (2018). Gender and Speech Disfluency Production: a Psycholinguistic Analysis on Turkish Speakers. In *Psycholinguistics*, *24* (2), 114-143. https://doi.org/10.31470/2309-1797-2018-24-1-114-143

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. In *Journal of Statistical Software*, 67 (1), 1-48. https://doi.org/10.18637/jss.v067.i01

Betz, S. (2020). Hesitations in Spoken Dialogue Systems. (Doctoral dissertation). Universität Bielefeld.

Betz, S., Eklund, R. & Wagner, P. (2017). Prolongation in German. In *Proceedings of DiSS 2017 The 8th Workshop on Disfluency in Spontaneous Speech*, Stockholm, Sweden, 18-19 August 2017, 13-16.

Betz, S., Wagner, P. (2016). Disfluent lengthening in spontaneous speech. In *Elektronische Sprachsignalverarbeitung (ESSV) 2016*, Leipzig, Germany, 2-4 March 2016.

Boersma, P., Weenink, D. (2021). Praat: Doing phonetics by computer [computer program]

Crocco, C., Savy, R. (2003). Fenomeni di esitazione e dintorni: una rassegna bibliografica. In Crocco, C., Savy, R. & Cutugno, F. (a cura di), *API. Archivio di Parlato Italiano*, DVD.

Den, Y. (2003). Some strategies in prolonging speech segments in spontaneous speech. In Eklund, R. (Ed.), *Proceedings of the ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech (DiSS 2003)*, Göteborg, Sweden, 5-8 September 2003, 87-90.

Di Napoli, J. (2020). Filled pauses and prolongations in Roman Italian task-oriented dialogue. In *Proceedings of the Laughter and Other Non-Verbal Vocalisations Workshop*, Biefeld, 5 October 2020, 24-27.

Eklund, R. (2001). Prolongations: A dark horse in the disfluency stable. In *Proceedings of DiSS 2001 Disfluency in Spontaneous Speech*, Edinburgh, Scotland, UK, 29-31 August 2001, 5-8.

Eklund, R. (2004). Disfluency in Swedish human-human and human-machine travel booking dialogues. PhD Dissertation, Linköping University: Electronic Press.

Eklund, R., Shriberg, E. (1998). Crosslinguistic disfluency modelling: a comparative analysis of Swedish and American English human-human and human-machine dialogues. In *Proc. 5th International Conference on Spoken Language Processing (ICSLP 1998)*. https://doi.org/10.21437/ICSLP.1998-756

Gafos, A., Van Lieshout, P. (2020). Models and theories of speech production. In *Frontiers in Psychology*, 11, 1238.

Giannini, A. (2003). Hesitation phenomena in spontaneous Italian. In *Proceedings of the 15th International Congress of Phonetic Sciences* (ICPhS), 2653-2656.

Ginzburg, J., Fernández, R. & Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. In *Semantics and Pragmatics*, 7 (9), 1-64. https://doi.org/10.3765/sp.7.9

Gósy, M., Eklund, R. (2017). Segment prolongation in Hungarian. The 8th Workshop on Disfluency in Spontaneous Speech (DiSS 2017), 29–32.

Gósy, M., Eklund, R. (2018). Language-Specific Patterns of Segment Prolongation in Hungarian. *In Phonetician*, *115*, 36-52.

Hamzah, R., Jamil, N. & Seman, N. (2012). Acoustical analysis of filled pause in Malay spontaneous speech. In *Computer Applications for Communication, Networking, and Digital Contents: International Conferences, FGCN and DCA 2012, Held as Part of the Future Generation Information Technology Conference, FGIT 2012, Gangneug, Korea, December 16-19, 2012. Proceedings* (pp. 251-259). Springer Berlin Heidelberg.

Hope, R.M. (2012). Rmisc: *Ryan Miscellaneous*. R package version 1.5.1, https://cran.r-project.org/web/packages/Rmisc

Jabeen, F., Wagner, P. (2023). Variability in hesitations in Punjabi semi-spontaneous narrative speech: An automatic clustering based analysis, In *Disfluency in Spontaneous Speech (DiSS) Workshop 2023*, ISCA, 71–75. https://doi.org/10.21437/DiSS.2023-15

Johnson, W. (1961). Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers. *Journal of Speech and Hearing Disorders. Monograph Supplement*, 1–20.

Kosmala, L. (2024). *Beyond disfluency: The interplay of speech, gesture, and interaction*. John Benjamins.

Kuruoglu, G., Altiparmak, A. (2015). A Psycholinguistic point of view on prolongations of the native speakers of Turkish. In *International Journal of Arts & Sciences*, *8*(3), 185.

Landis, J.R., Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Lee, T.-L., He, Y.-F., Huang, Y.-J., Tseng, S.-C. & Eklund, R. (2004). Prolongation in spontaneous Mandarin. In *Proceedings of Interspeech 2004*, Jeju Island, Korea, 4-8 October 2004, 2181-2184.

Lenth, R., Singmann, H., Love, J., Buerkner, P. & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. R package version, 1 (1), 1–97.

Levelt, W.J. (1983). Monitoring and self-repair in speech. In *Cognition*, 14(1), 41-104.

Levelt, W.J. (1989). *Speaking: From intention to articulation* (Vol. 1). Cambridge, MA: MIT Press.

Li, X., Ishi, C.T., Fu, C. & Hayashi, R. (2022). Prosodic and Voice Quality Analyses of Filled Pauses in Japanese Spontaneous Conversation by Chinese learners and Japanese Native Speakers. In *Proceedings of Speech Prosody 2022*, 550–554. https://doi.org/10.21437/SpeechProsody.2022-112

Lickley, R.J. (2017). Disfluency in typical and stuttered speech. In Bertini, C., Celata, C., Lenoci, G., Meluzzi, C. & Ricci, I. (Eds.), *Fattori sociali e biologici nella variazione fonetica*. Studi AISV 3, 373–387. Milano: Officinaventuno. https://doi.org/10.17469/O2103AISV000019

Lickley, R.J. (2015). Fluency and Disfluency. In *Redford, M.A. (Ed.), The handbook of speech production*. Chichester: John Wiley & Sons, 445-474. https://doi.org/10.1002/9781118584156.ch20

Lobanov, B. (1971). Classification of Russian vowels spoken by different listeners. In *Journal of the Acoustical Society of America*. 49. 606–608.

Maekawa, K., Mori, H. (2017). Comparison of Voice Quality between the Vowels in Filled Pauses and Ordinary Lexical Items. In *Journal of the Phonetic Society of Japan*, vol. 21, no. 3, 53-62.

Mayer, M. (1969). *Frog, ¿where are you?*. New York: Dial Press.

Moniz, H., Mata, A.I. & Céuviana, M.C. (2007). On filled-pauses and prolongations in European Portuguese. In *8th Annual Conference of the International Speech Communication Association*. Antwerp, Belgium, 27-31 August, 2820-2824 https://doi.org/10.21437/Interspeech.2007-695

Onslow, M. (2023). Stuttering and its treatment: Twelve lectures. Last retrieved October 20, 2024, from https://www.uts.edu.au/asrc/resources

R core team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. https://www.R-project.org/

Sarro, G. (2023). 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.

Schettino, L. (2022). The Role of Disfluencies in Italian Discourse. Modelling and Speech Synthesis Applications. (Doctoral dissertation). Università degli Studi di Salerno.

Schettino, L., Betz, S., Cutugno, F. & Wagner, P. (2021a). Pitch and Functional Characterization of Hesitation Phenomena in Italian Discourse. In *Proceedings of the 4th Phonetics and Phonology in Europe* (PaPE 2021).

Schettino, L., Betz, S., Cutugno, F. & Wagner, P. (2021b). Hesitations and individual variability in Italian tourist guides' speech. In *Speaker individuality in phonetics and speech sciences: Speech technology and forensic applications*, Studi AISV 8, 243–262. Milano: Officinaventuno.

Schettino, L., Campisi, E. & Origlia, A. (2024) Disfluenze bimodali. Forme, funzioni e pattern di fenomeni di disfluenza e gesti per la gestione del parlato di guide turistiche. In Cirillo, L., Nodari, R. (a cura di) Studi AItLA 18: *Contesti, pratiche e risorse della comunicazione multimodale*, 209–223. Milano: Officinaventuno.

Schettino, L., Eklund, R. (2023). Prolongation in Italian. In *Disfluency in Spontaneous Speech (DiSS) Workshop 2023*, ISCA, 81–85. https://doi.org/10.21437/DiSS.2023-17

Shriberg, E.E. (1994). Preliminaries to a theory of speech disfluencies (Doctoral dissertation). University of California.

Silber-varod, V., Gósy, M. & Eklund, R. (2019). Segment prolongation in Hebrew, in *Proceedings of DiSS 2019*. The 9th Workshop on Disfluency in Spontaneous Speech, ELTE Faculty of Humanities, 47–50. https://doi.org/10.21862/diss-09-013-silb-etal

Sloetjes, H., Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakesh, Marocco, 28-30 May 2008, 816-820.

Stepanova, S. (2007). Some features of filled hesitation pauses in spontaneous Russian. In: *Proceedings of ICPhS*.

Voghera, M., Mayhofer, V., Ricci, M., Rosi, F. & Sammarco, C. (2020). Il Modokit: un identikit modale delle produzioni parlate e scritte dalle medie al biennio. In Voghera, M., Maturi, P. & Rosi, F. (Eds.), *Orale e scritto, verbale e non verbale: la multimodalità nell'ora di lezione*, 227-245. Firenze: Cesati.

Voghera, M. (2017). *Dal parlato alla grammatica*. Roma: Carocci.

Williams, S. (2022). Disfluency and proficiency in second language speech production. Cham: Palgrave Macmillan.

Wingate, M.E. (1984). Fluency, disfluency, dysfluency, and stuttering. In *Journal of Fluency Disorders*, 9 (2), 163–168.

Zmarich, C. (2017). Stuttering and phonetic theory: An introduction. In Bertini, C., Celata, C., Lenoci, G., Meluzzi, C. & Ricci, I. (Eds.), *Social and biological factors in speech variation*, Studi AISV 3, 357–372. Milano: Officinaventuno.