Disfluencies in continuous speech in French: Prosodic parameters of filled pauses and vowel lengthening

Yaru Wu[1,2], *Ivana Didirková*[3], *Anne-Catherine Simon*[4]

[1]EA 4255 CRISCO, Université de Caen Normandie, France
[2]UMR 7018 Laboratoire de Phonétique et Phonologie (CNRS Sorbonne-Nouvelle), France
[3]UR 1569 TransCrit, Université Paris 8 Vincennes - Saint-Denis, France
[4]Centre VALIBEL, Institut Langage et Communication, Université catholique de Louvain, Belgium

yaru.wu@unicaen.fr, ivana.didirkova@univ-paris8.fr, anne-catherine.simon@uclouvain.be

Abstract

This study examines prosodic parameters in two types of disfluencies, vowel lengthenings, and filled pauses. We analyzed approximately 2.5 hours of continuous speech representing 11 different speech genres, including prepared, semi-prepared, and unprepared speech. Mean fundamental frequency (f0) and pitch resets were analyzed to compare the prosodic correlates of disfluent (e.g., lengthenings, filled pauses) syllables with their surrounding fluent syllables as a function of the degree of speech preparation. The results show that the average fundamental frequency is lower in filled pauses and disfluent vowel lengthenings than in fluent speech. Furthermore, filled pauses are produced with a lower f0 than vowel lengthening. Larger pitch resets are observed between disfluent units and their preceding contexts, and both filled pauses and prolongations depend on the degree of preparation of the discourse. The duration of vowel lengthening tends to be longer than that of filled pauses.
Index Terms: disfluency, vowel lengthening, filled pause, pitch, duration

1. Introduction

The phenomenon of disfluency has long been studied in the language sciences. Including filled pauses (*euh* in French), vowel lengthening, false starts, repetitions, verbal productions indicating a correction, or silent pauses ([1]), disfluencies are often interpreted as a sign of difficulty in the speech production process. As such, their frequency should decrease with the opportunity to plan or rehearse one's speech. Moreover, disfluencies are also presented as indicative of the unpredictability of speech ([2]) and can be perceptually exploited as the announcement of new or complex content ([3, 4]). Thus, while they are often presented as characteristic of (unprepared) spontaneous speech ([5]), disfluencies are also regularly produced in formal or prepared styles ([6]).

From a prosodic point of view, previous studies have shown a tendency for the fundamental frequency to decrease locally on the disfluent syllable or region (-0.9 ST in Portuguese, [5]) and to increase between the disfluent and the immediately following region (+0.8 ST). Using a corpus of eight languages, [7] showed that filled pauses (or autonomous vowels of hesitation support) mostly have a duration between 200 ms and 650 ms. Regarding fundamental frequency, the authors found no significant difference between hesitations and fluent vowels (schwa and [ø]). However, hesitant vowels often had an irregular vocal quality. In French, the fundamental frequency value of filled pauses is similar to the onset value of breath groups for a given speaker ([8]). [4] studied filled pauses and vowel lengthening in French and found that the mean duration of vowel lengthening (347.25 ms) was significantly higher than the mean duration of filled pauses (268.4 ms). The distribution of vowel lengthening durations is also narrower than that of filled pauses ([4]).

2. Aim and hypotheses

We focus on the prosodic correlates of two similar disfluency markers: hesitation vowel lengthening (abnormal syllable lengthening at the beginning or end of a word with a flat or slightly descending intonational contour ([1])) and filled pauses (explicit markers of quasi-lexicalized hesitation, usually transcribed as *euh* in French, and representing and epenthesis at the end of a word or pronounced independently ([1])) produced in different speech situations. More specifically, we aim to provide mean fundamental frequency values, pitch reset, and mean duration values for hesitation vowel lengthening, filled pauses, and fluent syllables as a function of speech preparation level. Our hypotheses are as follows: 1. We expect disfluency duration to be shorter in prepared speech compared to unprepared speech; 2. We expect pitch reset to be more pronounced between filled pauses and their preceding syllable compared to hesitation lengthening and the preceding syllable, but comparable pitch reset between these two disfluency types and the following syllable; 3. We expect the fundamental frequency on disfluent syllables to decrease less in prepared speech than in unprepared speech.

3. Method

3.1. Corpus

Our analyses are conducted on the LOCAS-F (Louvain Corpus of Annotated Speech - French) corpus ([9]), a multigenre oral corpus annotated for disfluencies, syntactic units, and prosodic units. The corpus contains 42 sound samples representing 14 different speech genres (or communicative activities). In this study, samples containing conversational narratives had to be discarded due to numerous overlaps that could have distorted the f0 readings. Our analyses thus cover 2h38min35sec of recordings (average recording length: 4min53) and include the following genres and degrees of preparation: 1. prepared speech (the whole intervention is written, then read in front of the audience): read speech, academic speech, sermon, radio news, political speech; 2. semi-prepared speech (the content is prepared but the speech remains improvised to some degree): radio narrative, conference; 3. unprepared speech (the text was not prepared and is comparable to everyday spontaneous interactions): interview (formal, free, radio), conversational storytelling.

3.2. Manual data annotation

Disfluencies were annotated according to a protocol adapted to verbal and signed languages ([10]): simple disfluencies, based on tokenization of the orthographic transcription, and compound disfluencies, grouping multiple tokens ([4]). Our analyses cover more than 1500 vowel prolongation sequences, 1000 filled pauses, and 59000 fluent syllables. Disfluency thus affects slightly more than 4 % of the data, which is a rather low proportion in non-pathological speech. The annotation of filled pauses and hesitation lengthening was performed manually by experts and evaluated on a part of the corpus by calculating an agreement rate from the annotations of the three experts ([4]). For filled pauses, the kappa value is 0.86 (Z = 85.55, near perfect agreement), and for hesitation lengthening, 0.64 (Z = 63.11, substantial agreement). Pre-final boundary lengthening was considered phonological and, therefore, fluent.

3.3. Data selection

The analyses focus on three target categories. Two of the three categories are related to disfluency (i.e., vowel lengthening (LG) and filled pause (FP)), and they are compared to disfluency-free speech (Fluent). Please note that the category "Fluent" contains a few other disfluencies (less than 1%) that we did not cover in this paper. Measurements were carried out in Praat ([11]) with an ad-hoc script. Pitch reset was calculated (1) by subtracting the fundamental frequency of the last syllable of the preceding word from the fundamental frequency of the fluent sequence, the filled pause, or the lengthened vowel (= `studied sequence (Fluent/LG/FP) − last syllable of the previous word`) and (2) subtracting the fundamental frequency of the analyzed sequence from the fundamental frequency of the first syllable of the following word (= `first syllable of the following word − studied sequence (Fluent/LG/FP)`). The fundamental frequency was converted to relative semitones (ST) at 50 Hz. Data are split into three categories depending on the preparation degree of the speech (see above).

4. Analyses and results

Linear mixed models (LMM) were used to analyze the average fundamental frequency and pitch reset as a function of our target variable "F_DF" (consisting of three levels: 1) fluent; 2) vowel lengthening; 3) filled pause) using the *lme4* package in R ([12]). Similarly, an LMM model was run for the duration of vowel lengthening and filled pause. Four different models were implemented because four different data sets were involved. For the first three models (one for the average fundamental frequency and two for the pitch reset on both sides of the studied sequence), "F_DF" was included as a fixed effect. Preparation level (i.e., prepared, semi-prepared, and unprepared speech) and speaker gender were included as control variables. For random effects, intercepts were included for subject and item for the first three models on (1) mean f0, (2) pitch reset between "F_DF" and the last syllable of the preceding word, (3) pitch reset between the first syllable of the following word and "F_DF". In the last model on the duration of filled pause and vowel lengthening, "LG/FP" and "degree of preparation" were included as fixed effects. Speaker Gender was included as a control variable. The intercept was included for subject and phoneme. For statistical analyses, duration was z-normalized across recording sessions.

4.1. Average fundamental frequency

Figure 1 shows the mean values of fundamental frequency in fluent sequences compared to vowel lengthening and filled pauses for both female (on the left) and male (on the right) speakers. The results show that the mean fundamental frequency is lower in vowel lengthening and filled pauses (mean = 18.04 ST, standard deviation (SD) = 6.70 ST) than in fluent speech (mean = 19.01 ST, SD = 5.97 ST). This observation is confirmed for both females (with mean values of 22.30 ST (SD = 4.96 ST) in disfluent and 23.66 ST (SD = 4.30 ST) in fluent speech) and males (15.42 ST (SD = 6.28 ST) in disfluent and 16.72 ST (SD = 5.32 ST) in fluent speech), which represents a decrease of more than 1 ST in disfluent sequences. If we decompose disfluencies into vowel lengthening and filled pauses, we observe the same trend with a more pronounced decrease of f0 in filled pauses (mean = 17.90 ST, SD = 7.31 ST) than in vowel lengthening (mean = 18.10 ST, SD = 6.43 ST) and than in fluent sequences (mean = 19.01 ST, SD = 5.97 ST). This

difference is observed both in females with an average f0 of 21.54 ST (SD = 5.26 ST) in filled pauses and 22.71 ST (SD = 4.73 ST) in vowel lengthening, and in males whose average f0 is 14.97 ST (SD = 7.41 ST) in filled pauses and 15.59 ST (SD = 5.81 ST) in vowel lengthening. The decrease in f0 in filled pauses reaches 1.75 ST in men and 2.12 ST in women compared to fluent speech. During lengthening, this decrease is more limited, about 1 ST (0.95 ST in women and 1.14 ST in men).

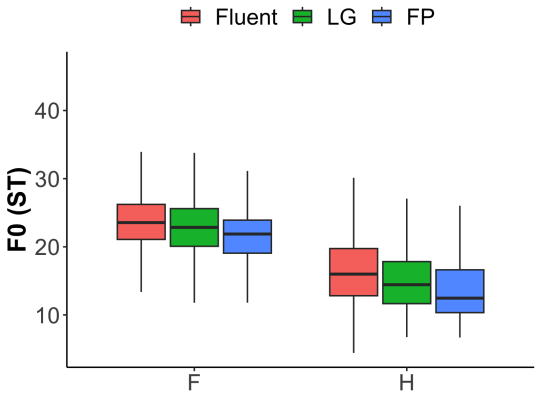

Figure 1: *F0 average in fluent vs. disfluent sequences and as a function of gender. Fluent sequences in red ("Fluent"), vowel lengthening ("LG") in green, and filled pause ("FP") in blue. F = female, H = male.*

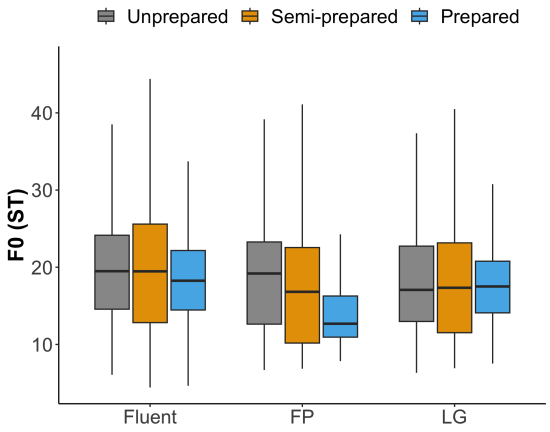

Figure 2: *Average f0 according to the degree of preparation and the type of disfluency. In grey, unprepared (NP), in orange semi-prepared (SP) and in blue prepared (P) speech.*

Finally, by examining the average f0 as a function of the degree of preparation, it appears that the behavior of filled pauses in prepared speech differs from that of other degrees of preparation. The average f0 of filled pauses decreases with the degree of preparation: the more prepared the speech, the lower the average f0 of these disfluencies. In unprepared speech, their average f0 is 18.89 ST (SD = 7.11 ST). In semi-prepared speech it is 17.41 ST (SD = 7.91 ST) and in prepared speech 14.24 ST (SD = 5.01 ST), i.e. a decrease of 4.66 ST between unprepared and prepared speech. Note that vowel lengthening does not follow this trend: with values of 18.21 ST (SD = 6.42 ST) and 18.14 ST (SD = 5.60 ST), respectively, the average f0 difference between unprepared and prepared speech in these disfluencies is 0.07 ST. Finally, in fluent speech, a decrease of 1.2 ST is observed between unprepared (mean = 19.68 ST, SD = 6.23 ST) and prepared (mean = 18.48 ST, SD = 5.29 ST) speech, as shown in Figure 2. The LMM results confirm that the mean f0 of "LG" [β = -0.48307; t = -20.831; SE = 0.02319] and "FP" [β = -1.56979; t = -17.372; SE = 0.09036] is significantly lower than that of "Fluent". The model-based post-hoc test shows significant differences between Fluent, LG and FP ($p < 0.001$) for the dependent variable (f0). No significant difference was found for the mean of f0 as a function of the level of preparation.

4.2. Pitch reset

Pitch reset results are presented for melodic changes between (1) the last syllable of the previous word and the studied sequence (i.e., the fluent sequence ("Fluent"), the vowel lengthening ("LG") or the filled pause ("FP")) on the one hand, and (2) the studied sequence ("Fluent," "LG" or "FP") and the first syllable of the next word on the other hand (see the details of the computations in Section 3.3). Disfluencies are not shown in the figure for the prepared speech of females, since there are only a few occurrences.

Figure 3 shows that for the pitch reset between the last syllable of the preceding word and the vowel lengthening ("LG"), the filled pause ("FP"), or the fluent sequence ("Fluent"), lengthened vowels and filled pauses generally produce a greater negative melodic change than fluent sequences, except for the visually undetectable difference between Fluent and LG for semi-prepared speech in females. Results based on the corresponding LMM model confirm that LG [β = -0.42520; t = -2.520; SE = 0.16871] and FP [β = -2.59138; t = -6.376; SE = 0.40645] elicit more negative melodic changes than Fluent. The post-hoc test based on the LMM model shows that "Fluent", "LG" and "FP" are significantly different from each other ($p < 0.001$ for all comparisons). We also examined the pitch reset between the fluent sequence, vowel lengthening, filled pause, and the first syllable of the following word. However, no significant difference was found for the pitch reset between the studied sequence and the first syllable of the following word according to the LMM model.

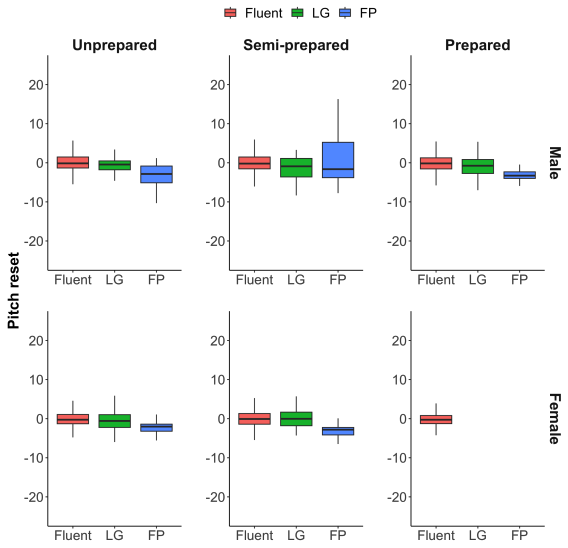

Figure 3: *The pitch reset (in ST) between the last syllable of the preceding word and the fluent sequence ("Fluent", in red), the vowel lengthening ("LG", in green), or the filled pause ("FP", in blue), as a function of the degree of preparation and gender (from left to right: unprepared, semi-prepared, and prepared utterances; top part for males, bottom part for females)*

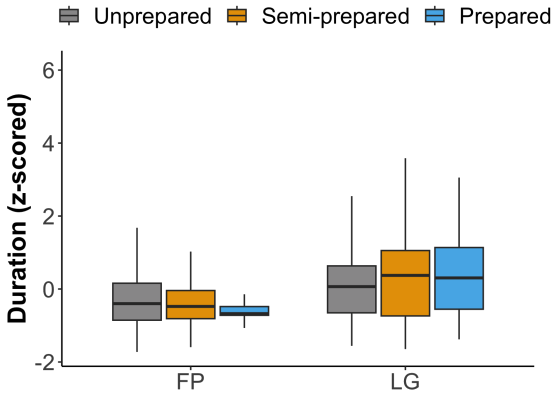

Figure 4: *Normalized duration of filled pauses ("FP") and vowel lengthenings ("LG") according to the degree of preparation (from left to right: unprepared, semi-prepared, and prepared speech).*

4.3. Duration of filled pauses and vowel lengthening

Figure 4 shows the normalized duration of filled pauses ("FP") and vowel lengthenings ("LG") as a function of the degree of preparation. Interestingly, opposite patterns are observed for the duration of filled pauses ("FP") and vowel lengthening ("LG"). The more prepared the speech style, the shorter the filled pauses. As for vowel lengthening ("LG"), shorter durations are found in unprepared speech than in semi-prepared and prepared speech. The LMM results confirm that "FP" tend to have a significantly shorter duration than "LG" [β = -0.74633; t = 7.95; SE = 0.09389]. No significant difference was found for the level of preparation.

5. Discussion

The aim of this study was to describe the melodic behavior of lengthened vowels and filled pauses in comparison with fluent sequences in French, in speech styles with different degrees of preparation. Indeed, we assumed that the ability to prepare speech could influence disfluency behavior. Specifically, we examined the average fundamental frequency in disfluent and fluent sequences as a function of the degree of speech preparation, the pitch reset that separates disfluent from fluent sequences, and the duration in these two sequences.

Our results show a decrease in fundamental frequency in disfluent sequences, in line with the results obtained for Portuguese ([5]) and American English ([13]). However, there is a difference between vowel prolongations and filled pauses, the former being less pronounced in terms of fundamental frequency. This difference, which is also observed in Hungarian ([14]), could be explained by the nature of disfluency itself, with filled pauses forming autonomous items. Furthermore, filled pauses are also characterized by a greater drop in f0 during prepared speech. Another parameter would also be the habit of working with one's own voice, especially for broadcast news journalists, who represent a third of the corpus of prepared speech.

This result is also supported by the one obtained on pitch reset, where filled pauses and vowel lengthening stand out from the preceding syllable by decreasing f0. Interestingly, filled pauses also decrease the fundamental frequency of the following syllable, which is not the case for vowel lengthening.

Finally, we observed longer durations for lengthened vowels compared to filled pauses, confirming the results in [4]. Filled pauses were longer in unprepared speech than in prepared speech. This result is not surprising and corresponds to the frequency of these events in unprepared and prepared speech.

Our analyses should be completed with other parameters, especially the location of the disfluency, both in grammatical terms (filled pause placed between two different syntactic units or within a unit) and in discourse terms, taking into account the prosodic environment, which would allow us to better explain their specific behavior in terms of vowel lengthening. Furthermore, other disfluencies (mainly repetitions and false starts) should be studied in order to refine our conclusions.

6. References

[1] M. Candea, "Contribution à l'étude des pauses silencieuses et des phénomènes "d'hésitation" en

français oral spontané. Etude sur un corpus de récits en classe de français," Ph.D. dissertation, Université Paris III-La Sorbonne Nouvelle, 2000.

[2] F. Goldman-Eisler, *Psycholinguistics: experiments in spontaneous speech*, academic press ed., London, New York, 1968.

[3] M. Watanabe, K. Hirose, Y. Den, and N. Minematsu, "Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners," *Speech Communication*, vol. 50, no. 2, pp. 81–94, 2008.

[4] I. Grosman, "Évaluation contextuelle de la (dis)fluence en production et perception : pratiques communicatives et formes prosodico-syntaxiques en français," Ph.D. dissertation, UCL - Université Catholique de Louvain, 2018. [Online]. Available: https://dial.uclouvain.be/pr/boreal/object/boreal:208866

[5] H. Moniz, A. I. Mata, and C. Viana, "On filled-pauses and prolongations in european portuguese," in *Proceedings of Interspeech 2007*. Anvers, Belgique: ISCA, 2007, pp. 2645–2648.

[6] A.-C. Simon, A. Auchlin, M. Avanzi, and J.-P. Goldman, "Les phonostyles: une description prosodique des styles de parole en français," vol. 2, pp. 71–88, 2010.

[7] I. Vasilescu, M. Candea, and M. Adda-Decker, "Hésitations autonomes dans 8 langues: une étude acoustique et perceptive," in *Actes du colloque MIDL*, Paris, 2004, pp. 25–30.

[8] D. Duez, "Acoustic-phonetic characteristics of filled pauses in spontaneous French speech: preliminary results," in *Proceedings of DiSS'01, Disfluency in Spontaneous Speech Workshop*. Edinburg, UK: ISCA, 2001, pp. 41–44.

[9] L. Martin, L. Degand, and A. C. Simon, "LOCAS-F : un corpus oral multigenres annoté," in *CMLF 2014 - 4 ème Congrès Mondial de Linguistique Française*. Berlin: EDP Sciences, 2014, pp. 2613–2626.

[10] L. Crible, A. Dumont, I. Grosman, and Notarrigo, "Annotation manual of fluency and disfluency markers in multilingual, multimodal, native and learner corpora. Version 2.0," Université catholique de Louvain, Tech. Rep., 2016.

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

[12] R Development Core Team, *R: A La nguage and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[13] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, pp. 521–524 vol.1.

[14] A. Deme and A. Markó, "Lengthenings and filled pauses in hungarian adults' and children's speech," *Proceedings of Disfluency in Spontaneous Speech, DISS 2013*, vol. 54, p. 21, 2013.