Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts

Helena Moniz¹,², Isabel Trancoso², Ana Isabel Mata¹

¹FLUL/CLUL, University of Lisbon, Lisbon, Portugal ²IST / INESC-ID, Lisbon, Portugal

helenam@l2f.inesc-id.pt, isabel.trancoso@inesc-id.pt, aim@fl.ul.pt

Abstract

This work explores prosodic cues of disfluent phenomena. In our previous work, we conducted a perceptual experiment regarding (dis)fluency ratings. Results suggested that some disfluencies may be considered felicitous by listeners, namely filled pauses and prolongations. In an attempt to discriminate which linguistic features are more salient in the classification of disfluencies as either fluent or disfluent phenomena, we used CART techniques on a corpus of 3.5 hours of spontaneous and prepared non-scripted speech. CART results pointed out 2 splits: break indices and contour shape. The first split indicates that events uttered at breaks 3 and 4 are considered felicitous. The second shows that these events must have flat or ascending contours to be considered as such; otherwise they are strongly penalized. Our preliminary results suggest that there are regular trends in the production of these events, namely, prosodic phrasing and contour shape.

Index Terms: prosody, disfluency, fluency rating

1. Introduction

Disfluencies, e.g., filled pauses, prolongations, repetitions, substitutions, deletions, insertions, characterize spontaneous speech and play a major role in speech structuring [1][2][3]. For speech processing, the analysis of the regular patterns of those phenomena is crucial [4][5]. In automatic speech recognition, their identification accounts for more robust language and acoustic models [6] and even in speech synthesis, these phenomena are being modeled to improve the naturalness of synthetic speech [7].

The fluent component of those phenomena is still rather controversial, even though [8][9] have already pointed out the benefits of disfluencies for communicative purposes, and their contribution for on-line planning efforts. Moreover, the crosslinguistic properties of those events, mainly filled pauses, show regular trends[10][11], pointing out linguistic principles and parameters. Taking these claims into account, can we say that all disfluencies behave alike? What linguistic features play a major role in the production of disfluencies? Are they really disfluent when they have a pragmatic and metalinguistic function? Can we delete them all in order to obtain the intended message, as in a scripted version of speech?

Preliminary studies for European Portuguese (e.g., [12][13]) have mainly targeted filled pauses and segmental prolongations. Those studies suggested that the regular patterns observed in the production and perception of these specific types of disfluencies are related to different levels of the prosodic structure. They also claimed that, due to their prosodic specificities, they may behave as fluent devices.

From a production perspective, different filled pauses tend to occur in different prosodic contexts: (i) aam generally occurs at major intonational phrase boundaries, (ii) aa is most likely found at minor intonational phrase boundaries; (iii) mn is criticized onto prior elongated words. Segmental prolongations are more likely found at internal clause boundaries, and at a constituent level, behaving as aa. The studies also pointed out that filled pauses are uttered mainly with stationary contours, whereas segmental prolongations exhibit more complex F0 contours.

From a perception point of view, these studies wanted to test if all types of disfluencies should be rated as infelicitous, or conversely, if disfluencies in different prosodic contexts, and with different contour shapes could be rated as felicitous or infelicitous. This was the motivation for conducting a perceptual test in which 40 participants classified a number of stimuli as felicitous and infelicitous moments concerning ease of expression in a 5-point scale. When only stimuli whose average score was above or equal 4 were considered felicitous, three different sets of disfluency phenomena emerged, which are associated with different acceptability rates: (1) prolongations and filled pauses; (2) substitutions and deletions; (3) segments, repetitions and complex sequences. Prolongations were better rated than filled pauses, and repetitions were strongly penalized. Prolongations and filled pauses rated as felicitous moments were regularly scaled relatively to their adjacent constituents, a behavior that did not stand for filled pauses and repetitions occurring in infelicitous moments.

Silent pauses are consistently used as a cue to either automatically recognize disfluencies [14] or to analyse their psycholinguistic implications [1][3]. Our previous study [15] pointed out that more than 80% of prolongations and filled pauses are followed by silent pauses of a reasonable length, supporting the view that their presence may effectively be used by listeners as a cue to an upcoming delay. The absence of such a pause is strongly penalized as misleading information.

As for phrasing, the existence of an intermediate phrase level [16] across languages and for a specific language is still a matter of debate. This prosodic constituent corresponds to a break index 3 in the ToBI system [17]. In the joint attempt to propose a ToBI system for European Portuguese [18], the authors working with professional reading and spontaneous speech data pointed out the importance of having the break index 3 for speech processing. This level could account for sentence-like chunks, the description of disfluencies, and the way they relate to adjacent prosodic constituents.

We now aim at validating the assumption that prosodic phrasing is crucial to perform a fluency/disfluency rating task, using Classification and Regression Trees techniques (CART)
Our concrete goal in this work is to find out what linguistic features are more salient when we classify all types of disfluencies as either fluent or disfluent phenomena. This task is harder than it seems, since fluency is a complex notion, and not even expert annotators can objectively state that the prosodic behavior is more salient than the morphosyntactic or semantic ones. Although the bulk of the paper is devoted to our CART experiment and its relationship with the perceptual experiments, the next section will briefly describe the corpora used in this work.

2. Corpora

This work uses subsets of the CPE-FACES [20] and LECTRA [21] corpora. Whereas the first corpus includes spontaneous and prepared non-scripted speech at high-school (two teachers and twenty five students), totalling 15h, the second one includes university presentations (five teachers), totalling 10h. Subsets of these corpora were manually annotated for disfluencies and fluency ratings: 2h, for the high school corpus, and 1.5h for the university one. The disfluency tier was annotated according to [5] and [22]. Additional tiers were added with prosodic (break indices, contour shape and F0 restart) and part of speech information (POS of the disfluency and adjacent words). The information from the different annotation tiers was organized into a database and an annotator added the perceptual judgments of the disfluencies, i.e., whether the uttered events were fluent or disfluent.

The disfluency rate is 13.24% (1569 disfluencies and 11,851 words) in the high school corpus, and 3.16% (273 disfluencies and 8636 words) in the university corpus. A randomly selected sample of the first corpus was also annotated by two other expert linguists, in terms of ease of expression, as felicitous or infelicitous. The agreement between the three annotators was of 95%.

3. CART Experiment

Our CART experiment was conducted using the SAS software1. We started by dividing the annotated data of the two corpora into training, validation and test data (60%, 20% and 20%, respectively). The test misclassification rate was 29.05%. The features used were: (dis)fluent judgements (as target feature), disfluency type, break indices, F0 contour, F0 restart, morphosyntactic information of the adjacent words, morphosyntactic information of the disfluency, speaker and speech situation (spontaneous and prepared non-scripted speech).

The results shown in Figure 1 indicate that 56.4% of the events are classified by the CART as disfluent and 43.6% as fluent. The first split in the tree is on the variable break indices. This variable allows the distinction between disfluencies uttered within a prosodic constituent (classified most often as infelicitous), and at break indices 3 and 4 (classified as felicitous). Within a constituent, 78.3% of these events are infelicitous, and the remaining 21.7% are considered fluent.

4. Relationship with perceptual test

The above results are consistent with the findings of the perceptual test [13] that had also pointed out the importance of break indices and phrasing in fluency judgements. This motivated a detailed study of all the prosodic constituents of the stimuli.

Our study targeted three types of disfluencies: segmental prolongations, filled pauses and repetitions. These specific types of disfluencies have been considered by [23][3] as associated to planning efforts. In corpora of school presentations and lectures, which are intrinsically associated with clarifying messages and planning carefully what to say next, these types of disfluencies are thus worth studying in detail.

Figure 2 represents a stylization in semitones (ST) of the disfluency and its prosodic context. For each one, we have plotted the maximum and the offset of the previous constituent; the onset, maximum and offset of the disfluent event; and the onset and maximum of the subsequent prosodic constituent. The F0 measurements are not represented in the real temporal intervals.

As the figure shows, prolongations judged felicitous exhibit F0 ascending contours with high sustained boundary tones, typically observed at the end of a prosodic constituent with continuation meaning. Filled pauses also judged fluent are uttered in a tonal space in between the prosodic adjacent constituents.
have stationary F0 contours and behave mostly as parenthetica-
ments. When filled pauses are considered infelicitous, however,
they are produced in a lower register with descending contours,
disrupting tonal scaling. As for repetitions, the examples that
we have tested were prosodically illformed and considered dis-
fluent (e.g., lexical and function words repeated), we did not in-
clude emphatic repetitions or rheortical ones. The disfluent rep-
etitions behave mostly as disfluent filled pauses, but were pre-
ceded by strong melodic disruptions.

Figure 3 represents a felicitous example of a prolongation 
\[s'\text{er} \alpha\] (ser, ‘to be’) uttered at a break 3 with an ascending F0
contour and a high boundary tone with continuation meaning.
This high boundary tone is realized in the appended elongated
vowel \(x\). The prolongation is adequately adjusted to the ad-
joined prosodic constituents and scaled relatively to the adjacent
F0 peaks.

Figure 3: Felicitous example: “passa a ser um homem aven-
tureiro” (‘he becomes an adventurous man’)

An example judged disfluent is illustrated in figure 4, where
the verb [s'\text{er} \alpha\] (s\(\text{ao} \), ‘are’) is repeated. As in the first example,
the repetition by itself forms a prosodic constituent, in this spe-
cific case with a descending contour. The unit disrupts the F0
global contour, and consequently the scaling between peaks.

The results of Figure 2 partially agree with the ones of [24]
and [25], in the way that filled pauses have linear and gradual
F0 descending contours. However, in our data, they may exhibit
ascending or flat contours as well. As pointed out by [25], filled
pauses tend to be uttered between the previous peak and the
baseline of the speaker. A result that was also observed in our
data is that these events are uttered at a tonal space in between
adjacent prosodic constituents.

The segmental and suprasegmental characteristics of dis-
fluencies, some idiomsincratic and some general, may be seen as
contributions for the discussion of their classification as regular
words, and for the possible delimitation of these phenomena as
a minor prosodic constituent when they do not coarticulate with
the previous word. These important phonetic and prosodic cues
used by listeners to signal planning efforts at different levels of
the prosodic structure, and these levels may be
identified with different pragmatic functions.

In previous work, we have pointed out that segmental pro-
longations did not seem to undergo regular sandhi processes for
European Portuguese. For instance, the adversative conjunction
\textit{mas} (‘but’) when the vowel \[x\] is appended is often pronounced
as [\textmu x\textit{e}z\textmu] instead of [\textmu x\textmu z\textmu]. For filled pauses produced within
a prosodic constituent, these findings seem to hold as well for
our present data. Examples such as \textit{efetos especiais} \(a\) (‘spe-
cial effects uh’), with no silent pause or glottalization interval
between the second word and the filled pause, are pronounced
as [\textmu j\textit{ij}j\textmu t\textmu z\textmu ]\(a\)\(po\)\(s\)\(a\)\(y\)\(\textmu t\textmu n\) instead of [\textmu z\textmu]. The regular sandhi
process in European Portuguese is applied in the coarticulation
of the two words but not between the last one and the filled
pause [\textmu n\textmu t\textmu n\textmu ].

5. Conclusions and future work

Previous sections have shown that prosodic phrasing is cru-
ical to perform an evaluation task regarding fluency/disfluency
distinctions, but contour shape also plays an important role in
this kind of task. Both the CART and the perceptual exper-
iments pointed out that disfluencies may behave and even be
rated as fluent devices. Results suggest, in line with findings
for other languages, that speakers control different segmental
and suprasegmental aspects, and they seem to do it, in many
cases, in a \textit{surgical} way - adequately adjusting to the adjacent
constituents.

This work may be seen as a rehearsal for a more detailed
study, aiming at the discrimination between fluent and disfluent
phenomena in spontaneous and prepared non-scripted speech,
based on larger corpora for European Portuguese.

The fluent component of these communicative devices
poses a terminology problem - should we continue to call them
disfluencies, a term widely used by the scientific community,
when they behave fluently in certain contexts? We would like
to extend our work and try to answer this question. Another
research direction is to analyse all the segmental and supraseg-
mental cues of all types of these events in order to automati-
cally identify them. We also want to analyze the different prag-
matic functions they may have. For instances, in our current
data, filled pauses precedes new information or computer jar-
gon translations into European Portuguese. The communicative
devices we have been describing seem, thus, to be uttered at dif-
frent levels of the prosodic structure, and these levels may be
associated with different pragmatic functions.

6. Acknowledgements

The authors would like to thank Tiago Luis for is technical
support and our two annotators who classified the university
lecture and provide us with useful comments. This work was
funded by FCT SFRH/BD/44671/2008, and FCT projects
INESC-ID Lisboa had support from the POSI Program of the
“Quadro Comunitário de Apoio III”.

7. References

1989.


ing,” Cognition, no. 84, 2002.

cues in spontaneous speech,” Journal of the Acoustical Society of
America (JASA), no. 95, pp. 1603–1616, 1994.

Figure 4: *Infelicitous example: “a música o ballet e a dança moderna são são os principais da cultura cubana”* (‘music, ballet, and modern dance are are the principal [aspects] of cubane culture’)


