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

Helena Moniz[1,2], *Isabel Trancoso*[2], *Ana Isabel Mata*[1]

[1]FLUL/CLUL, University of Lisbon, Lisbon, Portugal
[2]IST / 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) *mm* is cliticized 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 contrarily, 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) fragments, 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 repetions 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)

[19]. 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 software[1]. 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 (spontaneos 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 classified as fluent devices. The latter (21.7%) are uttered either at the onset of an intonational phrase and have F0 restart (10.4%), or at the end of a constituent with boundary tones that signal continuation (break 3) or finality (break 4), as in neutral statements in European Portuguese.

The second split in the tree (F0 contours) shows that events produced at breaks 3 or 4 with flat or ascending contours are mainly considered fluent (90%) vs. the ones uttered in similar positions, but with descendent contours or with glottalization effects (72%).

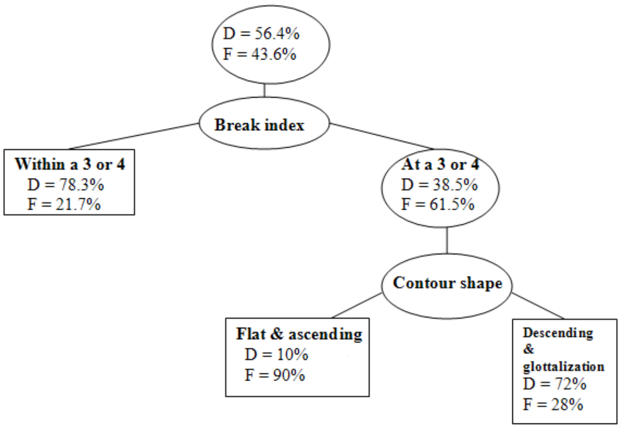

Figure 1: *CART results. “D” stands for disfluent/infelicitous, and “F” for fluent/felicitous classification.*

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.

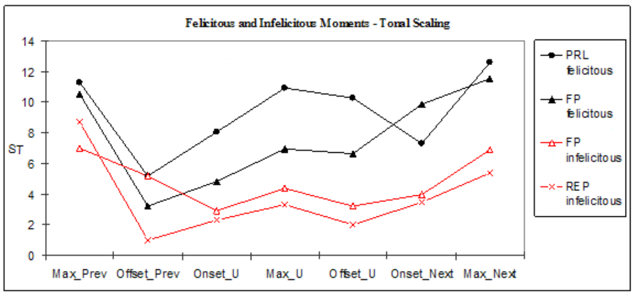

Figure 2: *Tonal scaling of prolongations, filled pauses and repetitions judged (in)felicitous.*

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,

[1] http://www.sas.com

have stationary F0 contours and behave mostly as parentheticals. 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 disfluent (e.g., lexical and function words repeted), we did not include emphatic repetitions or rethorical ones. The disfluent repetitions behave mostly as disfluent filled pauses, but were preceded by strong melodic disruptures.

Figure 3[2] represents a felicitous example of a prolongation [s'erəː] (*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 [əː]. The prolongation is adequately adjusted to the adjoined prosodic constituents and scaled relatively to the adjacent F0 peaks.

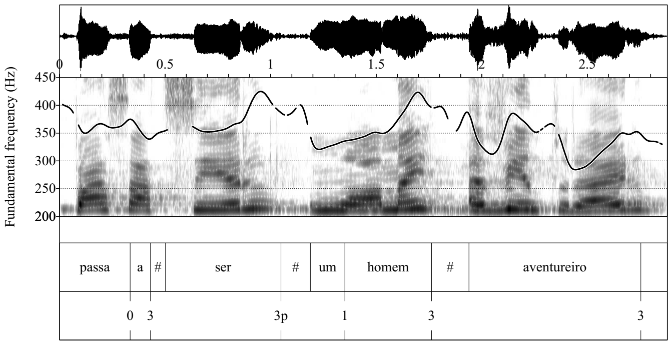

Figure 3: *Felicitous example: "passa a ser um homem aventureiro"('he becames an adventurous man')*

An example judged disfluent is illustrated in figure 4, where the verb [s'ɒw̃̃] (*são*, 'are') is repeated. As in the first example, the repetition by itself forms a prosodic constituent, in this specific 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 disfluencies, some idiossincratic 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 may be used to identify disfluencies in automatic speech recognition applications.

In previous work, we have pointed out that segmental prolongations did not seem to undergo regular *sandhi* processes for European Portuguese. For instance, the adversative conjunction *mas* ('but') when the vowel [əː] is appended is often pronounced as [mɒʒəː] instead of [mɒzəː]. For filled pauses produced within a prosodic constituent, these findings seem to hold as well for our present data. Examples such as *efeitos especiais aa* ('special effects uh'), with no silent pause or glottalization interval between the second word and the filled pause, are pronounced as [ʃ] [if'ɒjtuz əʃpəsj'ajʃ ɒː] instead of [z]. 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 [ɒː].

5. Conclusions and future work

Previous sections have shown that prosodic phrasing is crucial 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 experiments 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 *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 suprasegmental cues of all types of these events in order to automatically identify them. We also want to analyze the different pragmatic functions they may have. For instances, in our current data, filled pauses precedes new information or computer jargon translations into European Portuguese. The communicative devices we have been describing seem, thus, to be uttered at different levels of the prosodic structure, and these levels may be associated with different pragmatic functions.

6. Acknowledgements

The authors would like to thank Tiago Luís 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 PTDC/PLP/72404/2006 and CMU-PT/HuMach/0039/2008. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

7. References

[1] W. Levelt, *Speaking*. Cambridge, Massachusetts: MIT Press, 1989.

[2] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of Pragmatics*, no. 30, pp. 485–496, 1998.

[3] H. Clark and J. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, no. 84, 2002.

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

[5] E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California, 1994.

[2] figure done with praat and Pauline Welby's scripts.

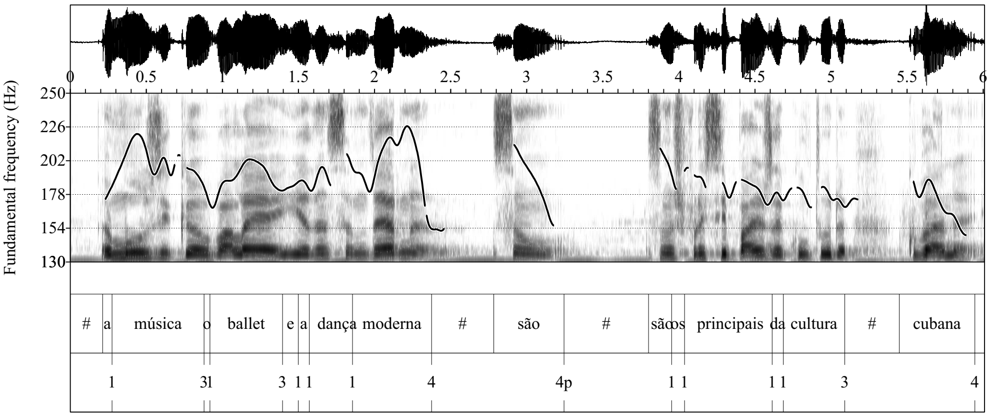

Figure 4: *Infelicitous example:"a música o ballet e a dança moderna são são os principais da cultura cubana" ('music, ballet, and modern dance are are the principal [aspects] of cubane culture')*

[6] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transaction on Audio, Speech, and Language Processing*, no. 14, pp. 1526–1540, 2006.

[7] J. Adell, A. Bonafonte, and D. Escudero-Mancebo, "On the generation of synthetic disfluent speech: local prosodic modifications caused by the insertion of editing terms," in *Proc. Interspeech '2008*, Brisbane, Australia, Sep. 2008.

[8] A. Heike, "A content-processing view of hesitation phenomena," *Language and Speech*, no. 24, 1981.

[9] D. O'Connell and S. Kowal, "Uh and um revisited: are they interjections for signaling delay?" *Psycholinguistic Research*, no. 34, 2005.

[10] R. Eklund and E. Shriberg, "Crosslinguistic disfluency modeling: a comparative analysis of swedish and american english human-human and human-machine dialogs," in *International Conference on Spoken Language Processing*.

[11] I. Vasilescu and M. Adda-decker, "A cross-language study of acoustic and prosodic characteristics of vocalic hesitations," in *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue*, A. E. M. B. E. Keller and M. Marinaro, Eds. IOS Press, 2007, pp. 140–148.

[12] H. Moniz, "Contributo para a caracterização dos mecanismos de (dis)fluência no Português Europeu," 2006.

[13] H. Moniz, A. I. Mata, and M. C. Viana, "On filled pauses and prolongations in European Portuguese," in *Proc. Interspeech '2007*, Antwerp, Belgium, Sep. 2007.

[14] A. Stolcke, E. Shriberg, T. Bates, M. Ostendorf, D. Hakkani, M. Plauché, G. Tür, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words," in *International Conference on Spoken Language Processing*, no. 5, 1999, pp. 2247–2250.

[15] H. Moniz, A. I. Mata, I. Trancoso, and M. C. Viana, "How can we use disfluencies and still sound as a good speaker?" in *Proc. Interspeech '2008*, Brisbane, Australia, Sep. 2008.

[16] M. Beckman and J. Pierrehumbert, "Intonational structure in japanese and english," *Phonology Yearbook*, no. III, pp. 15–70, 1986.

[17] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: a standard for labeling english prosody," in *International Conference on Spoken Language Processing*, Banff, Canada, 1992.

[18] M. C. Viana, S. Frota, coordenators, I. Falé, I. Mascarenhas, A. I. Mata, H. Moniz, and M. Vigário, "Towards a p_tobi," in *Unpublished Workshop of the Transcription of Intonation in the Ibero-Romance Languages, PaPI 2007*, http://www2.ilch.uminho.pt/eventos/PaPI2007/Extended-Abstract-P-ToBI.PDF, 2007.

[19] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Pacific Grove, Ca: Wadsworth and Brooks, 1984.

[20] A. I. Mata, "Para o estudo da entoação em fala espontânea e preparada no Português Europeu," Ph.D. dissertation, University of Lisbon, 1999.

[21] I. Trancoso, R. Martins, H. Moniz, A. I. Mata, and C. Viana, "The lectra corpus - classroom lecture transcriptions in european portuguese," in *LREC 2008 - Language Resources and Evaluation Conference*, Marrakesh, Morocco, May 2008.

[22] R. Eklund, "Disfluency in Swedish human-human and human-machine travel booking dialogues," Ph.D. dissertation, University of Linkopink, 2004.

[23] H. Clark and T. Wasow, "Repeating words in spontaneous speech," *Cognitive Psychology*, no. 37, 1998.

[24] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *IEEE Conference on Acoustic, Speech, and Signal Processing*, 1992, pp. 521–524.

[25] E. Shriberg, "Phonetic consequences of speech disfluency," in *International Congress of Phonetic Sciences*, San Francisco, 1999, pp. 612–622.