Prosodic context-based analysis of disfluencies

Helena Moniz[1,2], *Fernando Batista*[1,3], *Isabel Trancoso*[1,4], *Ana Isabel Mata*[2]

[1]INESC-ID, Lisbon, Portugal
[2]FLUL/CLUL, University of Lisbon, Portugal
[3]ISCTE-IUL - Lisbon University Institute, Lisbon, Portugal
[4]IST, Lisbon, Portugal
{helenam;fmmb;isabel.trancoso}@l2f.inesc-id.pt and aim@fl.ul.pt

Abstract

This work explores prosodic cues of disfluencies in a corpus of university lectures. Results show three significant ($p < 0.001$) trends: pitch and energy slopes are significantly different between the disfluency and the onset of fluency; those features are also relevant to disfluency type differentiation; and they do not seem to be a speaker-effect. The best combination of linguistic features one can use to better predict the onset of fluency are pitch and energy resets as well as the presence of a silent pause immediately before a repair. Our results, thus, point out to a strategy of prosodic contrast rather than of parallelism. With this work we hope to contribute to the analysis of the prosodic behaviors in the production of the so called disfluencies and in the fluency repair in European Portuguese.

Index Terms: prosody, disfluencies, university lectures.

1. Introduction

This paper describes the analysis of the prosodic properties of disfluencies and of their contexts, aiming both at a view of their global properties, and also at an analysis of their idiosyncratic behaviors. Disfluencies, *e.g.*, filled pauses, prolongations, repetitions, substitutions, deletions, characterize spontaneous speech and play a major role in speech structuring [1, 2]. For speech processing, the analysis of the regular patterns of those phenomena is crucial [3, 4, 5]. In automatic speech recognition (ASR), their identification accounts for more robust language and acoustic models [5] and even in text to speech synthesis (TTS), they are being modeled to improve the naturalness of synthetic speech [6]. Moreover, when combining ASR and TTS with speech-to-speech translation systems, spontaneous speech translation still needs substantial improvements [7]. Recently, disfluencies have also been studied as a feature in the detection of social behaviors [8].

We aim at analyzing: i) if there are different prosodic cues for distinct types of disfluencies and ii) if there are correlations between the prosodic properties of the disfluencies and those of their adjacent contexts. The answer to these questions will hopefully be a step forward in two directions: contributing to a characterization of the so called disfluencies and of the fluency repair in European Portuguese (EP), based on empirical evidence supporting linguistic regularities at different levels, and, consequently, building predictive models based on regular trends in the prosodic behavior of the disfluencies and of their contexts.

2. Related work

In a disfluent sequence there are several regions to be considered: the *reparandum* or the region to be repaired, the *interregnum* interval, and the repair itself [9, 3, 4]. The possible connections between the *reparandum* and the repair have been explored with different perspectives in the literature. Since [9] there is a binary tendency towards the classification of the prosodic properties of (certain) disfluencies as either copying the pitch contour of the *reparandum* or contrasting the onset of fluency in the repair with the *reparandum*, by means of increasing f_0 and energy. The first strategy is classified as a parallelism between the two regions and is mainly related to appropriateness (involving, for instance, repetition and insertion), whereas the second is classified as contrast marking and is productive with error corrections (mostly substitutions). The literature is not consensual about this dichotomy. For [10], repetitions *per se* can behave as parallelistic prosodic structures (copying the pitch contour of the *reparandum*) and also have some degree of contrast (a rising pattern in the repetition is related to an emphasis in the new unit), although not the one reported by [9]. For [11], distinct categories, such as repetitions and substitutions seem to copy the patterns of their counterparts in the *reparandum*. Moreover, this study also shows that there is only partial support for the contrastive nature of substitutions when this is manifested by a higher pitch range. [12] sustains the parallelistic nature of both repetitions and error corrections and considers parallelism the most frequent strategy.

For European Portuguese, much has been said for silent pauses, filled pauses and prolongations (*e.g.*, [13, 14]), whereas the other categories are poorly described (with the exception of [15, 16]). We know that different filled pauses tend to occur in different prosodic contexts (*e.g.*, *aam* at major intonational phrase boundaries and *mm* in coda position). Segmental prolongations are more likely found at internal clause boundaries and at intonational phrase boundaries. In [15, 16], prosodic properties, mainly prosodic phrasing and contour shape, of all types of disfluencies are studied in their relation with an evaluation task regarding fluency/disfluency distinctions. The main results reported are that disfluencies may behave and even be rated as fluent communicative devices, when different segmental and suprasegmental aspects are monitored.

3. Corpus

The corpus collected within the *Lectra* project [17] aimed at transcribing university lectures for e-learning applications. The corpus has a total of 75h, corresponding to 7 different courses, of which 27h were orthographically transcribed. The speech recognition system [18] was used to produce the force aligned transcription. The reference data was then provided to the aligned transcription using the NIST SCLite tool. The corpus was automatically annotated with part-of-speech information using MARv [19]. The subset used in this work is the train subset. Table 1 presents its overall characteristics. The total alignment error for this subset was 10.2%, comparing relatively well with [20] for the same domain, although the amount of data is significantly different. The percentage of disfluencies is in line with findings by [21], who reported an interval of 5% to 10% in human-human conversation.

Disfluencies were annotated accordingly to [4] and [22]. Two or more consecutive disfluencies of the same nature were discriminated from a single one, in order to check if their prosodic behavior would be distinct. This resulted in 13 categories of disfluencies, as shown in Table 2. A single filled pause and a complex sequence are the most frequent types. Two or more consecutive deletions are the only category that is more frequent than a single item.

Speaker	Time (h)	Words	% Disfls	WER
S1	2.35	11790	2.93%	6.4
S2	2.00	9975	3.16%	21.5
S3	3.26	21296	**1.79%**	7.3
S4	2.35	12663	3.46%	25.2
S5	1.36	12366	5.38%	0.1
S6	4.05	29758	3.02%	7.3
S7	1.42	12579	**6.22%**	3.6
Total	**16.79**	**110427**	**3.46%**	**10.2**

Table 1: Overall characteristics of the training subset.

Type	S1	S2	S3	S4	S5	S6	S7	Total
Complex	115	100	116	124	124	247	227	**1053**
Deletion	4	8	22	14	5	46	14	**113**
Deletions	9	12	44	60	1	45	21	**192**
Filled pause	92	51	41	60	375	160	323	**1102**
Filled pauses	0	1	4	6	4	3	19	**37**
Fragment	13	42	31	20	11	55	30	**202**
Fragments	0	0	4	1	0	4	3	**12**
Prolongation	34	0	0	0	45	30	17	**126**
Prolongations	1	0	0	0	1	2	5	**9**
Repetition	25	44	41	54	50	172	62	**448**
Repetitions	20	12	18	54	8	33	22	**167**
Substitution	20	35	46	27	24	79	25	**256**
Substitutions	9	8	14	18	9	19	9	**86**
Total	**342**	**313**	**381**	**438**	**657**	**895**	**777**	**3803**

Table 2: Distribution of disfluencies per speaker. "S" stands for speaker.

4. Prosodic analysis

Pitch and energy were extracted using the Snack Sound Toolkit[1]. Durations of phones, words, and interword-pauses were extracted from the recognizer output. A set of syllabification rules was designed for Portuguese and applied to the lexicon. Features were calculated for the disfluent sequence itself and also for the two contiguous words, before and after the disfluent sequence. The following set of features has been used for each word in those regions: f_0 and energy raw and normalized mean, median, maxima, minima, and standard deviation, as well as POS, number of phones, and durations. Energy and f_0 slopes within the words were calculated based on linear regression.

4.1. Overall prosodic characterization

We first analyzed if there would be an overall tendency to both f_0 and energy resets in the repair region, when all the speakers and all the different types of disfluencies are accounted for. As Figure 1 shows, there are, in fact, pitch and energy increases from the disfluency region ("disf" or *reparandum*) to the repair of fluency ("disf+1"). We should add that disfluencies are followed by silent pauses 99% of the times in our data. Due to the fact that our

[1] http://www.speech.kth.se/snack/

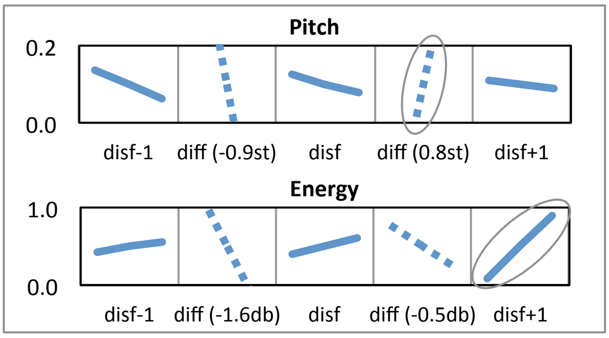

Figure 1: Pitch and energy slopes inside the disfluency (disf), word before (disf-1), and word after (disf+1); and differences between such units based on the average.

data set is nonparametric, we tested our hypotheses with a Kruskall-Wallis test. Results show significant differences with $p - value < 0.001$ in "disf-1", "disf" and "disf+1" pitch ($X^2_{12} = 53.82$; $X^2_{12} = 161.54$ and $X^2_{12} = 34.62$; respectively) and energy slopes ($X^2_{12} = 56.57$; $X^2_{12} = 78.09$ and $X^2_{12} = 152.47$; respectively) within a word as well as in the differences of pitch and energy amongst those regions (pitch and energy difference between "disf-1" and "disf"$X^2_{12} = 139.32$ and $X^2_{12} = 92.61$; between "disf" and "disf+1" $X^2_{12} = 378.34$ and $X^2_{12} = 104.95$; respectively). Thus, pitch and energy slopes are significantly different within the words immediately before and after the disfluencies (but not before and after that), meaning that contrasts are marked within relatively small contexts, possibly helping the listener to process useful cues in a shorter memory interval. These results have interesting implications for syntactic-prosodic mapping theories, supporting the view that a prosodic reset is an informative utterance suprasegmental planning cue [1].

4.2. Speaker and type of disfluency

Pitch and energy increase from the disfluency to the repair region, independently of the disfluency type and of the speaker (with the exception of speaker 3), as figures 2 and 3 show. There are, however, degrees in the pitch reset of the next unit. The highest pitch reset is after a filled pause (more than 2 ST) and it is significantly different ($p < 0.001$) from all the other disfluency types. This is, in fact, the disfluency with the subsequent prosodic context that most resemble the ones of a full-stop. Although filled pauses are the events that contribute the most to pitch increase, even without them pitch and energy resets are still significantly different ($X^2_{10} = 23.03$ with $p < 0.05$; ($X^2_{10} = 35.59$ with $p < 0.001$; respectively). We know that for EP ([13]), as for other languages, filled pauses tend to occur mainly at major intonational boundaries (*e.g.*, introducing topics), therefore the pitch and energy resets in the subsequent units are not that surprising. The second highest pitch reset occurs for the deletion type. Again, these findings are related to the fact that an unit after a deletion, as refreshed linguistic material, is more prone to exhibit an f_0 reset, which is an expected property at the beginning of a major intonational unit. As for energy, repetitions are significantly different ($p < 0.001$) from all the remaining types, with the highest energy slope within the repair. Even without repetitions, again pitch and energy resets are still significantly different ($X^2_{10} = 29.06$; $X^2_{10} = 42.33$; respectively).

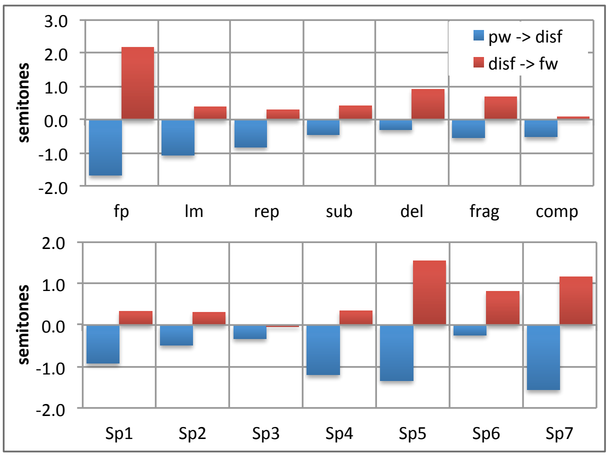

Figure 2: Difference between the disfluency, the previous and the following word pitch average, per type and speaker.

Additionaly, the prosodic contrast strategy does not apply exclusively to error correction categories (substitutions, deletions, fragments and complex sequences). Comparing with other types, substitutions, *e.g.*, show similar significant pitch/energy increase differences on the onset of the repair, or even on the slope within the repair. Thus, results do not support the use of a contrast strategy exclusively on the error corrections [9]. There is a more general tendency towards a contrast marking strategy, regardless of the specific disfluency type.

In what concerns intra-speaker variation, the global trends stand for the majority of the speakers. It is interesting to note that, when asked to classify the speakers regarding "likeability", our three annotators were unanimous in stating that speaker 6 is the most "likeable" one. The prosodic correlates of this naïve classification may be linked to several distinct linguistic features, namely, the highest energy slope within the repair, and also a considerable pitch increase, correlates which are frequently associated in fluent sequences with higher level strategies of language use and with charismatic speech [23].

5. Conclusions

The contributions of this work are threefold: firstly, pitch and energy slopes are significantly different within the units immediately before and after the disfluent se-

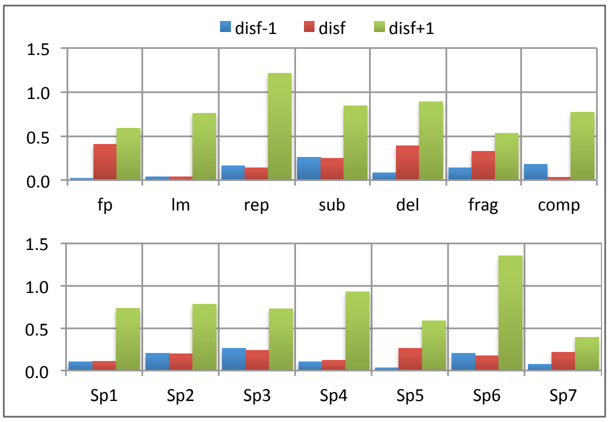

Figure 3: Energy slopes inside the previous word, disfluency, and in the following word, per type and speaker.

quences; secondly, pitch and energy increase in the repair in all disfluency types, but there are distinct degrees in the contrast made by certain ones (namely, filled pauses, deletions, and repetitions); and thirdly, those features are constant in the production of six of our seven speakers. These results can be discussed in different perspectives. Regarding the first contribution, the speaker signals the different regions, using the most economic way, just a word in the disfluency adjacent contexts, possibly helping the listener to process useful cues in a shorter memory interval. As for the second one, there are different contrastive degrees both in terms of pitch and energy (*e.g.*, filled pauses are the most distinct type in what regards pitch increase, and repetitions in what regards energy rising patterns). It seems, thus, that different prosodic parameters are combined in different degrees for different functional purposes. Finally, when repairing fluency, speakers overall produce both pitch and energy increases. Our analysis favors a tendency towards prosodic contrast strategies between the different regions. Thus, we could say that speakers signal different cues in sequences containing disfluencies, by means of contrast.

The economic and contrastive ways in which the speakers in our data manage speech may be associated with teaching skills, meaning that the speakers choose the most informative cues and yet less harmful from a listener comprehension point of view. Future work will tackle comparable studies for other domains, and also for other languages in the classroom domain.

6. Acknowledgments

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia – under Ph.D grant SFRH/BD/44671/2008 and projects CMU-PT/HuMach/0039/2008, PTDC/CLE-LIN/120017/2010, and by DCTI - ISCTE-IUL – Lisbon University Institute.

7. References

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

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

[3] 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.

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

[5] 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 Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[6] 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 *Interspeech 2008*, Brisbane, Australia, 2008.

[7] L. Tomokiyo, K. Peterson, A. Black, and K. Lenzo, "Intelligibility of machine translation output in speech synthesis," in *Interspeech 2006*, Pittsburgh, USA, September 2006.

[8] A. Gravano, R. Levitan, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic and prosodic correlates of social behavior," in *Interspeech 2011*, Florence, Italy, September 2011.

[9] W. Levelt and A. Cutler, "Prosodic marking in speech repair," *Journal of Semantics*, no. 2, 1983.

[10] M. Plauche and E. Shriberg, "Data-driven subclassification of disfluent repetitions based on prosodic features," in *ICPhS*, 1999.

[11] G. Savova and J. Bachenko, "Prosodic features of four types of disfluencies," in *DISS 2003*, 2003.

[12] J. Cole, J. Hasegawa-Johnson, C. Shih, H. Kim, E. Lee, H. Lu, Y. Mo, and T. Yoon, "Prosodic parallelism as a cue to repetition and error correction disfluency," in *DISS 2005*, 2005.

[13] H. Moniz, "Contributo para a caracterização dos mecanismos de (dis)fluência no Português Europeu," Master's thesis, University of Lisbon, 2006.

[14] A. Veiga, S. Candeias, C. Lopes, and F. Perdigão, "Characterization of hesitations using acoustic models," in *ICPhS 2011*, Hong Kong, China, August 2011.

[15] H. Moniz, I. Trancoso, and A. I. Mata, "Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts," in *Interspeech 2009*, Brighton, England, 2009.

[16] H. Moniz, I. Trancoso, and A. I. Mata, "Disfluencies and the perspective of prosodic fluency," in *Development of Multimodal Interfaces: Active Listening and Synchrony*, ser. LNCS, A. Esposito, N. Campbell, C. Vogel, A. Hussain, and A. Nijholt, Eds. Springer Berlin/ Heidelberg, 2010, vol. 5967, pp. 382–396.

[17] I. Trancoso, R. Martins, H. Moniz, A. I. Mata, and M. C. Viana, "The Lectra corpus - classroom lecture transcriptions in European Portuguese," in *LREC 2008 - Language Resources and Evaluation Conference*, Marrakesh, Morocco, May 2008.

[18] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in Portuguese," in *ICASSP 2008*, 2008, pp. 1561–1564.

[19] R. Ribeiro, "Anotação morfossintáctica desambiguada do português," Master's thesis, Instituto Superior Técnico, 2003.

[20] T. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in *Interspeech 2006*, Pittsburgh, U.S.A., September 2006.

[21] E. Shriberg, "To "errrr" is human: Ecology and acoustics of speech disfluencies," *Journal of the International Phonetic Association*, vol. 31, pp. 153–169, 2001.

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

[23] A. Rosenberg and J. Hirschberg, "Charisma perception from text and speech," *Speech Communication*, 2008.