Prosodic analysis of disfluent events in a corpus of university lectures

Helena Moniz[1,2], Ana Isabel Mata[2] & Isabel Trancoso[1,3]
[1] INESC-ID, [2] FLUL & [3] IST

This paper describes our efforts towards the analysis of the prosodic properties (pitch, energy, and duration) of disfluencies, aiming both at a view of their global properties, and also at an analysis of their idiosyncratic behaviors. Underlying this task is the fact that disfluencies, e.g., filled pauses, prolongations, repetitions, substitutions, deletions, insertions, characterize spontaneous speech and play a major role in speech structuring [l]. For speech processing, the analysis of the regular patterns of those phenomena is crucial [2,3]. In automatic speech recognition (ASR), their identification accounts for more robust language and acoustic models [4] and even in text to speech synthesis (TTS), they are being modeled to improve the naturalness of synthetic speech [5]. Moreover, when combining ASR and TTS with machine translation systems, spontaneous speech translation still needs substantial improvements [6].

In our previous work for European Portuguese [7], we proposed that prosodic properties, mainly prosodic phrasing and contour shape, are essential to perform an evaluation task regarding fluency/disfluency distinctions. In this perspective, disfluencies may behave and even be rated as fluent communicative devices, when different segmental and suprasegmental aspects are monitored. We now aim at extending our study to a characterization of the prosodic parameters of disfluencies in a more quantifiable way. Specifically, our main goal is to verify if disfluent events have distinct prosodic properties.

This work uses a subset of the LECTRA corpus [8], collected with the goal of transcribing university lectures for e-learning applications. The corpus has a total of 74h, of which 10h were multilayer annotated, including (besides other information) an orthographic tier, a morpho--syntactic tier, and a disfluency tier, annotated accordingly to [2]. This small subset corresponds to 5 speakers (the lecturers in each of the recorded courses). Table 1 shows the distribution of the different disfluencies for each of the speakers.

	al	eti	iou	oop	pmc	Total
filled pauses	163	98	589	164	246	**1260**
complex	123	70	107	132	214	**646**
repetitions	111	134	68	123	101	**502**
prolongations	70	32	109	125	101	**437**
deletions	55	105	18	76	47	**301**
substitutions	55	53	29	43	36	**214**
fragments	38	34	22	18	36	**148**
total	**613**	**526**	**942**	**681**	**746**	**3508**

Table 1 Distribution of disfluent types by speaker.

The f_0 and energy mean, maxima and minima, and total durations were automatically measured for each disfluent event[1]. A Kruskal-Wallis test shows significant differences ($p<0.001$) between all the events for each prosodic parameter analyzed. These findings support the view that at a global level the events are distinct. A search for the most significant differences amongst them shows that filled pauses, prolongations, and complex sequences are the most differentiable events in all the parameters.

1 f_0 slope is currently being considered as well.

Regarding f_0 parameters, f_0 *max* and f_0 *min* are more expressive than f_0 *mean* (which only allows for a distinction between filled pauses and prolongations *vs.* all the other events). Filled pauses are significantly lower than all the other types (for both f_0 *max* and f_0 *min*). Prolongations are also significantly different, since they are higher than filled pauses and lower than the remaining types. As for *energy* parameters, filled pauses are once more significantly different from all the other events, showing the highest energy *mean* and *min*, although they are not distinguishable from prolongations (z= -.847) with respect to *energy mean*. Complex sequences exhibit the highest *energy max* values. In what concerns the parameter *duration*, as expected, complex sequences and prolongations are the longest, being distinguishable from all the other events. Fragments, on the other hand, are significantly shorter than the others. There are only two disfluency types that are not significantly different when the seven parameters analyzed are taken into account: repetitions and substitutions.

Events	f_0 (ST)				Energy (dB)				Duration (ms)	
	median	mean	max	min	median	mean	max	min	median	mean
filled pauses	16.7	**16.6***	**19.4***	**13.4***	61.9	**61.2***	68.5	**45.2***	360	413
complex	18.7	18.9	**22.3***	**14.3***	59.6	**59.7***	**72.5***	41.5	600	**718***
Repetitions	19.1	18.9	21.3	15.5	58.3	58.2	69.4	42.7	300	383
prolongations	17.4	**17.8***	**20.2***	**14.1***	61.3	**60.9***	70.6	43.7	560	**619***
deletions	19.6	19.3	**22.4***	15.4	57.9	57.8	70.4	**40.3***	360	461
substitutions	19.0	18.7	21.0	15.7	57.9	57.9	68.9	43.1	320	343
fragments	19.0	19.1	21.6	16.4	56.9	56.1	68.1	42.6	250	**274***

Table 2 Prosodic properties of disfluent types. * stands for significant differences at the p < 0.001 level

We are currently exploring possible correlations between the prosodic properties of the disfluent events and the ones of their adjacent contexts, in order to better discriminate the idiosyncratic behaviors of disfluency types. This will allow us to understand how prosodic properties are monitored, and how their monitoring contributes to fluency/disfluency distinctions.

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

[2] Shriberg, E.: *Preliminaries to a Theory of Speech Disfluencies*. PhD. Thesis (1994).

[3] Nakatani, C., Hirschberg, J.: A corpus-based study of repair cues in spontaneous speech. *JASA*, (1994).

[4] Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., Harper, M.: Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transaction on Audio, Speech, and Language Processing* (2006).

[5] Adell, J., Bonafonte, A., Escudero-Mancebo, D: On the generation of synthetic disfluent speech: local prosodic modifications cased by the insertion of editing terms. *In Proc. Interspeech*, Brisbane, Australia, (2008).

[6] Tomokiyo, L., Peterson, K., Black, A., Lenzo, K.: Intelligibility of machine translation output in speech synthesis. *In: Proc. Interspeech*, Pittsburgh, USA, (2006).

[7] Moniz, H., Trancoso, I., Mata, A.I.: Disfluencies and the perspective of prosodic fluency. In Esposito, A. Et al., eds.: *Development of Multimodal Interfaces: Active Listening and Synchrony*, Springer-Verlag (2010).

[8] Trancoso, I., Martins, R., Moniz, H., Mata, A. I., Viana, M. C.: The Lectra corpus: classroom lecture transcriptions in European Portuguese. *LREC*, (2008).