Automatic structural metadata identification based on multilayer prosodic information

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

[1] Spoken Language Systems Lab – INESC-ID, Lisbon, Portugal
[2] FLUL/CLUL, Universidae de Lisboa, Portugal
[3] ISCTE – Instituto Universitário de Lisboa, Portugal
[4] IST, Lisboa, Portugal

Abstract

This paper discriminates different types of structural metadata in transcripts of university lectures: boundary events (comma, full stops and interrogatives), and disfluencies (repair). The disambiguation process is based on predefined multilayered linguistic information and on its hierarchical structure. Since boundary events may share similar linguistic properties, in terms of f_0 and energy slopes, presence/absence of silent pauses, and duration of different units of analysis, different classification methods based on a set of automatically derived prosodic features have been applied to differentiate between those events and disfluencies. This paper also performs a detailed analysis on the impact of each individual feature in discriminating each structural event. The results of our data-driven approach allow us to reach a structured set of basic features towards the disambiguation of metadata events. These results are a step forward towards the analysis of speech acts and their disambiguation from disfluencies.

Index Terms: disfluencies, automatic speech processing, structural metadata, speech prosody

1. Introduction

Enriching automatic speech transcripts with structural metadata [1, 2], namely punctuation marks and disfluencies, may highly contribute to the legibility of a string of words produced by a recognizer. This may be important for so many applications that the speech recognizer often appears integrated in a pipeline that also includes several other modules such as audio segmentation, capitalization, punctuation, and identification of disfluent regions. The task of enriching speech transcripts can be seen as a way to structure the string of words into several linguistic units, thus providing multilayered structured information which encompasses different modules of the grammar.

Different sources of information may be useful for this task, going much beyond the lexical cues derived from the speech transcripts, or the acoustic cues provided by the audio segmentation module (*e.g.,* speech/non-speech detection, background conditions classification, speaker diarization, etc.). In fact, one of the most important roles in the identification and evaluation of structured metadata is played by prosodic cues.

The goal of this paper is to study the impact of prosodic information in revealing structured metadata, addressing at the same time the task of recovering punctuation marks and the task of identifying disfluencies. The former is associated with the segmentation of the string of words into speech acts, and the later, besides other aspects, also allows the discrimination of potential ambiguous places for a punctuation mark. Punctuate spontaneous speech is in itself a quite complex task, further increased by the difficulty in segmenting disfluent sequences, and in differentiating between those structural metadata events. Annotators of the corpus used in this study report that those tasks are the hardest to accomplish, difficulty visible in the evaluation of manual transcripts, since the attribution of erroneous punctuation marks to delimit disfluent sequences corresponds to the majority of the errors. Furthermore, prosodic cues either for the attribution of a punctuation mark or for the signaling of a repair may be ambiguous [3, 4].

2. Related work

Recovering punctuation marks and disfluencies are two relevant MDA (Metadata Annotation) tasks. The impact of the methods and of the linguistic information on structural metadata tasks has been discussed in the literature. [5] report a general HMM (Hidden Markov Model) framework that allows the combination of lexical and prosodic clues for recovering *full stops*, *commas* and *question marks*. A similar approach was also used by [1, 6] for detecting sentence boundaries. A Maximum Entropy (ME) based method is described by [7] for inserting punctuation marks into spontaneous conversational speech, where the punctuation task is considered as a tagging task and words are tagged with the appropriate punctuation. It covers three punctuation marks: *commas*, *full stops*, and *question marks*; and the best results on the ASR output are achieved by combining lexical and prosodic features. A multi-pass linear fold algorithm for sentence boundary detection in spontaneous speech is proposed by [8], which uses prosodic features, focusing on the relation between sentence boundaries and break indices and duration, covering their local and global structural properties. Other recent studies have shown that the best performance for the punctuation task is achieved when prosodic, morphological and syntactic information are combined [1, 2, 9].

Much of the features and the methods used for sentence-like unit detection may be applied in disfluency detection tasks. What is specific of the latter is that disfluencies have an idiosyncratic structure: *reparandum*, interruption point, interregnum and repair of fluency [10, 11, 12]. The *reparandum* is the region to repair. The interruption point is the moment when the speaker stops his/her production to correct the linguistic material detected. Ultimately, it is the frontier between disfluent and fluent speech. The *interregnum* is an optional part and may include silent pauses, filled pauses (uh, um) or explicit editing expressions (I mean, no).

The repair is the corrected linguistic material. It is known that each of these regions has idiosyncratic acoustic properties that distinguish them from each other, inscribed in the edit signal theory [13], meaning that speakers signal an upcoming repair to their listeners. The edit signal is manifested by means of production of fragments, glottalizations, co-articulatory gestures and voice quality attributes, such as jitter (perturbations in the pitch period) in the *reparanda*.

Table 1: *Corpus properties and number of metadata events.*

Subset →	train+dev	test
Time (h)	28:00	3:24
number of words + filled pauses	216435	24516
number of disfluencies	8390	950
disfluencies followed by a repair	5608	720
number of full stops	8363	861
number of commas	22957	2612
number of question marks	3526	498

Sequentially, it is also edited by means of significantly different pause durations from fluent boundaries and by specific lexical items in the interregnum. Finally, it is edited via f_0 and energy contrastive or parallelistic patterns in the repair.

[14] present a statistical language model including the identification of POS tags, discourse markers, speech repairs, and intonational phrases, achieving better performances by analyzing those events simultaneously. Based on the edit signal theory, [11, 15] used CARTs to identify different prosodic features of the interruption point. [16, 1] used features based on previous studies and added language models to predict both prosodic and lexical features of sentence boundaries and disfluencies.

Our aim is to check if we are able to classify metadata structures based on the following set of features derived from the above-mentioned studies: pause at the boundary, pitch declination over the sentence, post-boundary pitch and energy resets, pre-boundary lengthening, word duration, silent pauses, filled pauses, and presence of fragments. By investigating how much one can classify and disambiguate in Portuguese, using just this set of very informative cues, we hope to contribute to the discussion of what are language and domain dependent effects in structural metadata evaluation.

3. Corpus

This study uses LECTRA[17], a corpus of university lectures transcribed for producing multimedia contents for e-learning applications. The corpus was divided into two main sets: *train+development* (89%), and *test* (11%). The sets include portions of each one of the courses and follow a temporal criterion, meaning the first classes of each course were included in the training set, whereas the final ones were integrated into both development and test sets. The data encompasses all the structural metadata events presented in Table 1.

One important aspect that characterizes Portuguese punctuation marks is the high frequency of *commas*, which in our corpus accounts for more than 50% of all events. In a previous study [3], where Portuguese and English Broadcast News are compared, the percentage of *commas* in the former is twice the frequency of the latter. The guidelines used for *commas* are the ones described in [18].

3.1. Integrating prosodic information

This work relies on information coming from the ASR output, manual transcripts, and the signal itself. After the speech recognition, all relevant manual annotations are transferred to the ASR transcripts, including all metadata events, by means of the NIST SCLite tool (http://www.nist.gov/speech). Durations of phones, words, and interword-pauses were extracted from the ASR output. Information regarding pitch (f_0) and energy (E) was not available in the ASR pipeline when this study started. For that reason, it has been directly extracted from the speech signal, using the Snack toolkit [19]. Energy and f_0 slopes were calculated based on linear regression.

Acoustic–phonetic parameters of segmental and supra-segmental units were, thus, automatically extracted to study structural metadata events. Organizing such information into hierarchies, meaning, into the smallest unit of analysis (phones or even sub-phone units) up to higher order constituents was crucial to the experiments conducted. At this point, the information extracted encompassed phones, syllables, words, sentence-like units, and speech-acts.

Our in-house speech recognizer [20], trained for the broadcast news domain, is totally unsuitable for the university lectures domain. The scarcity of text materials in Portuguese to train language models for this domain has motivated the decision of using the ASR in a force alignment mode, in order not to bias the study with the bad results obtained with an out-of-domain recognizer. For that reason, current experiments rely on force aligned transcripts that still contain about 0.9% of unaligned words (mainly due to low energy segments).

4. Predicting structural metadata events

Our experiments use a fixed set of purely automatic features, extracted either from the ASR output or from the speech signal itself. The features involve two words before the event and one word after the event, and characterize either a word or a sequence of two consecutive words. Features involving a single word include: pitch and energy slopes; ASR confidence score; word duration; number of syllables and number of phones. Features involving two consecutive words include: pitch and energy slopes shapes; pitch and energy differences; comparison of durations and silences before each word (*dur.comp*); and ratios for silences, word durations, pitch medians (*pmed.ratio*), and energy medians (*emed.ratio*). For example, *eslopes* : $RF_{cw,fw}$ is a shape feature that refers to the energy slope in the current (cw) and following words (fw), which is **R**ising in *cw* and is **F**alling in *fw*; $dur.ratio_{cw,fw}$ is a number between 0 and 1 that indicates the proportion of the duration of *cw* over the duration of *cw+fw*.

Our experiments were performed using the Weka toolkit [21] and distinct statistical methods have been applied, including: Naïve Bayes, Logistic Regression and Classification and Regression Trees (CART). The best results were consistently achieved using CARTs, closely followed by Logistic Regression. The remaining of this section shows the achieved results and performs an analysis on the most relevant features.

4.1. Results

Experiments aim at automatically detecting structural metadata events and at discriminating between those events, using mostly prosodic features (with the exception of two identical contiguous words). We have considered four different classes of structural elements, *full stop*s, *commas*, *question marks*, and disfluency repairs. Table 2 presents the best results achieved, using the standard metrics *precision, recall, F-measure* and *Slot Error Rate*. The best performance is achieved for *full stops,* confirming our expectation, since prosodic information is known to be crucial to classify those events in our language. The low results concerning *commas* are also justifiable, because our experiments rely on prosodic features, but *commas* depend mostly on lexical and syntactic features [9].

Table 2: *CART classification results for prosodic features.*

Class	**Precision**	**Recall**	**F-meas.**	**SER**
comma (,)	60.6	27.6	37.9	90.3
full stop (.)	64.1	67.6	65.8	70.2
question (?)	73.9	29.5	42.2	80.9
repair	60.8	13.1	21.6	95.4
weighted avg.	63.0	32.9	43.3	75.6

Table 3: *Confusion matrix between events.*

Classified as →	,	.	?	repair	*del.*
comma (,)	**718**	36	10	15	1823
full stop (.)	76	**579**	35	3	163
question (?)	27	225	**147**	4	95
repair	51	19	1	**93**	546
insertions	312	44	6	38	

The performance for *question marks* is mainly related to their lower frequency and to the multiple prosodic patterns found for these structures. Moreover, interrogatives in our language are not commonly produced with subject-auxiliary verb inversion, as in English, which renders the problem of identifying interrogatives even more challenging. The worse performance, specially affected by a low recall, is achieved for *repairs*. While prosodic features seem to be strong cues for detecting this class, the confusion matrix presented in Table 3 reveals that *repairs* are still confused with regular words. Table 3 also reveals that the most ambiguous class is, without doubt, interrogatives.

Our recent experiments as well as other reported work [10, 11, 12] suggest that *filled pauses* and *fragment*s serve as cues for detecting structural regions of a disfluent sequence. Supported by such facts, we have conducted an additional experiment using *filled pauses* and *fragments* as features. These features turned out to be amongst the most informative features, increasing the *repair* f-measure to 48.8%, and improving the overall f-measure to 47.8%. However, the impact of fragments is lower than the one reported by [11, 22] and this may be due to the fact that fragments in our corpus represent only 6.6% of all the disfluent types.

4.2. Most salient features

Equivalent experiments performed with Logistic Regression provide a good approximation to the impact of each feature. A first inspection on Table 4 suggests that two pairs of structural metadata events are prone to be classified as ambiguous: *full stops* and *question marks*; and *repairs* and *regular words*. However, a closer inspection reveals that a set of informative features stands out as determinant to disambiguate between such events, namely, pitch and energy shapes, duration ratios, and confidence levels of the units of analysis.

Features for the discrimination of a ***repair*** comprise: i) two identical contiguous words; ii) both energy and pitch increases in the following word and (mostly) a plateau contour on the preceding word; and iii) a higher confidence level for the following word than for the previous word. Reasoning about this, this set of features is showing that repetitions are being identified, that repair regions are characterized by prosodic contrast marking (increases in pitch and energy) between disfluency–fluency repair (as in our previous studies), and also that the repair identification has a high confidence level.

Table 4: *Top most relevant features, sorted by relevance.*

	Feature	**none**	**,**	**.**	**?**	**repair**
1	$pslopes : F\_{pw,cw}$			▪▪▪	▪▪▪▪▪	
2	$pslopes : --_{pw,cw}$			▪▪▪▪	▪▪▪▪▪	
3	$pslopes : R\_{pw,cw}$			▪▪▪▪	▪▪▪▪▪	
4	$conf_{cw}$			▪▪▪▪▪	▪▪▪▪	
5	$eslopes : RF_{cw,fw}$			▪▪▪▪▪	▪▪▪▪	
6	$eslopes : --_{pw,cw}$			▪▪▪▪▪	▪▪	
7	$eslopes : F\_{pw,cw}$			▪▪▪▪▪	▪▪▪	
8	$eslopes : R\_{cw,fw}$			▪▪▪▪▪	▪▪▪	
9	$eslopes : R\_{pw,cw}$			▪▪▪▪▪	▪▪	
10	$eslopes : RF_{pw,cw}$			▪▪▪	▪▪▪▪▪	
11	$eslopes : FF_{pw,cw}$			▪▪▪	▪▪▪▪▪	
12	$eslopes : RR_{cw,fw}$			▪▪▪▪	▪▪▪▪	
13	$eslopes : -F_{pw,cw}$			▪▪▪▪	▪▪▪▪	
14	$pslopes : RF_{cw,fw}$		▪	▪	▪▪▪▪▪	
15	$pslopes : F\_{cw,fw}$			▪▪▪▪▪	▪▪▪	
16	$pslopes : FF_{cw,fw}$			▪▪▪▪▪	▪▪▪	
17	$pslopes : R\_{cw,fw}$		▪	▪	▪▪▪▪▪	
18	$pslopes : RR_{cw,fw}$		▪	▪	▪▪▪▪▪	
19	$pslopes : FR_{cw,fw}$			▪▪▪▪▪	▪▪▪	
20	$bsil.ratio_{cw,fw}$	▪▪▪▪▪				
21	$bsil.comp :>_{cw,fw}$			▪▪▪	▪▪▪▪	
22	$emed.ratio_{cw,fw}$	▪	▪	▪▪▪	▪▪▪	
23	$bsil.ratio_{pw,cw}$	▪▪▪▪▪	▪▪			▪
24	$dur.ratio_{cw,fw}$		▪	▪▪▪▪▪		▪
25	$dur.ratio_{pw,cw}$	▪▪▪	▪▪	▪		▪
26	$emed.ratio_{pw,cw}$	▪▪	▪	▪▪▪		▪
27	$pslopes : -F_{pw,cw}$	▪▪▪▪	▪			▪▪▪
28	$pslopes : RF_{pw,cw}$	▪▪▪▪	▪▪			▪▪▪
29	$pslopes : FF_{pw,cw}$	▪▪▪▪	▪▪			▪▪▪
30	$pslopes : -F_{cw,fw}$	▪▪▪▪				▪▪▪▪▪
31	$eslopes : -F_{cw,fw}$	▪▪▪	▪▪			▪▪▪
32	$pslopes : --_{cw,fw}$	▪▪▪				▪▪▪▪▪
33	$equals_{pw,cw}$	▪	▪	▪▪▪		▪▪
34	$pslopes : -R_{cw,fw}$	▪▪▪				▪▪▪▪▪
35	$phones_{cw}$	▪	▪▪	▪▪	▪▪	▪
36	$bsil.comp :<_{cw,fw}$	▪▪▪	▪▪			▪▪
37	$bsil.comp :>_{pw,cw}$	▪	▪	▪▪	▪▪▪	▪
38	$eslopes : -R_{cw,fw}$	▪▪	▪▪			▪▪▪▪
39	$eslopes : --_{cw,fw}$	▪▪	▪▪			▪▪▪▪
40	$pmed.ratio_{pw,cw}$	▪▪	▪	▪▪	▪	▪
41	$eslopes : FR_{cw,fw}$		▪▪			▪▪▪▪▪
42	$pslopes : -R_{pw,cw}$	▪▪				▪▪▪▪▪
43	$eslopes : RR_{pw,cw}$	▪	▪			▪▪▪▪▪
44	$eslopes : -R_{pw,cw}$	▪	▪▪			▪▪▪▪▪
45	$eslopes : FR_{pw,cw}$	▪	▪▪			▪▪▪▪▪
46	$eslopes : F\_{cw,fw}$	▪	▪▪			▪▪▪▪▪
47	$pslopes : FR_{pw,cw}$	▪▪				▪▪▪▪▪
48	$pslopes : RR_{pw,cw}$	▪▪				▪▪▪▪▪
49	$equals_{cw,fw}$		▪		▪	▪▪▪▪▪
50	$eslopes : FF_{cw,fw}$	▪	▪▪			▪▪▪▪▪
51	$conf_{fw}$	▪	▪		▪	▪▪▪▪▪
52	$bsil.comp :=_{cw,fw}$	▪▪	▪	▪	▪	▪▪
53	$bsil.comp :=_{pw,cw}$	▪▪	▪▪	▪	▪	▪▪
54	$dur.comp :>_{cw,fw}$	▪	▪▪	▪	▪▪	▪
55	$dur.comp :<_{cw,fw}$	▪▪	▪	▪	▪	▪▪
56	$phones_{fw}$	▪	▪▪	▪▪	▪▪	▪
57	$pmed.ratio_{cw,fw}$	▪	▪	▪▪	▪▪	▪▪
58	$dur.comp :<_{pw,cw}$	▪▪	▪▪	▪▪	▪	▪
59	$dur.comp :>_{pw,cw}$	▪▪	▪	▪	▪▪	▪▪
60	$syls_{fw}$	▪▪	▪	▪	▪	▪▪

As for ***full stops***, the determinant prosodic features correspond to: i) a falling contour in the current word; ii) a plateau energy slope in the current word; iii) the duration ratio between the current and the following words; and iv) a higher confidence level for the current word.

Reasoning about this characterization, it is the one that most resemble the neutral statements in our language, with the canonical contour H+L* L%.

Question marks are characterized by two main patterns: i) a rising contour in the current word and a rising/rising energy slope between current and following words; and ii) a plateau pitch contour in the current word and a falling energy slope in the current word. The rising patterns associated with question marks are not surprising, since they commonly associated with interrogatives. The falling pitch contours have also been ascribed for different types of interrogatives, especially wh-questions in Portuguese.

Commas, as stated along the previous section, are the event characterized by fewer prosodic features. Being mostly identified by morphosyntactic features, they are not clearly disambiguated with prosodic features.

With regards to ***regular words***, the most salient features are related to the absence of silent pauses, explained by the fact that, contrarily to the other events, regular words within phrases are connected. The presence of a silent pause is a strong cue to the assignment of a structural metadata event.

The literature for Portuguese points out to an array of features relevant for the description of metadata events. With the exception of *commas*, the data-driven approach followed in this work allow us to reach a structured set of basic features towards the disambiguation of such events beyond the established evidences for language.

5. Conclusions

This paper reports experiments on a full discrimination of structural metadata events in a corpus of university lectures, a domain characterized by a high percentage of structural events, namely punctuation marks and disfluencies. Our previous work on automatic recovery of punctuation marks indicates that specific punctuation marks display different sets of linguistic features. This motivated the discrimination of the different SU types. Our experiments, purely based on prosodic features, achieved a considerable performance, further improved when the ground truth about filled pauses and fragments was also used. Moreover, based on a set of complex prosodic features, we were able to point out regular sets of features associated with the discrimination of events (*repairs*, *full stops*, and *question marks*).

Future experiments will extend this study to fully automatic speech recognition transcripts and evaluate how the discrimination between the punctuation marks and disfluencies is affected by the ASR errors. Future work will also tackle the inclusion of lexical and morphosyntactic features, which are expected to considerably improve the performance, specially for *commas* and *question marks*.

6. Acknowledgements

This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia, under Ph.D grant SFRH/BD/44671/2008, projects PEst-OE/EEI/LA0021/2013 and PTDC/CLE-LIN/120017/2010, and by ISCTE–IUL.

7. References

[1] 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 Trans. on Audio, Speech and Lang. Proc.*, vol. 14, no. 5, pp. 1526–1540, 2006.

[2] M. Ostendorf and et al, "Speech segmentation and spoken document processing", *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 59–69, 2008.

[3] F. Batista, H. Moniz, I. Trancoso, and N. Mamede, "Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts", *Transactions on Audio Speech and Language Processing*, no. 20, pp. 474–485, 2012.

[4] H. Moniz, F. Batista, I. Trancoso, and A. I. Mata, "Prosodic context-based analysis of disfluencies", in *Proc. Interspeech*, Portland, Oregon, 2012.

[5] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models", in *ASRU*, 2001.

[6] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics", *Speech Communication*, vol. 32, pp. 127–154, 2000.

[7] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," *in Proc. ICSLP*, 2002.

[8] D. Wang and S. S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues", in *Proc. ICASSP*, vol. 1, 2004.

[9] B. Favre, D. Hakkani-Tür, and E. Shriberg, "Syntactically-informed Models for Comma Prediction", in *ICASSP*, 2009.

[10] W. Levelt, Speaking. Cambridge: MIT Press, 1989.

[11] 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.

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

[13] D. Hindle, "Deterministic parsing of syntactic non-fluencies", in *Proc. ACL*, 1983.

[14] P. Heeman and J. Allen, "Speech repairs, intonational phrases and discourse markers: Modeling speakers' utterances in spoken dialogue", *Computational Linguistics*, vol. 25, pp. 527–571, 1999.

[15] E. Shriberg, "Phonetic consequences of speech disfluency", in *Proc. ICPhS*, San Francisco, 1999.

[16] J.-H. Kim and P. C. Woodland, "Automatic capitalisation generation for speech input", *Computer Speech & Language*, vol. 18, no. 1, pp. 67–90, 2004.

[17] I. Trancoso, R. Martins, H. Moniz, A. I. Mata, and M. C. Viana, "The Lectra corpus – classroom lecture transcriptions in European Portuguese", in *Proc. LREC*, Morocco, 2008.

[18] I. Duarte, Língua Portuguesa, Instrumentos de Análise. Lisboa: Universidade Aberta, 2000.

[19] K. Sjölander, J. Beskow, J. Gustafson, E. Lewin, R. Carlson, and B. Granström, "Web-based educational tools for speech technology", in *Proc. ICSLP*, Australia, 1998.

[20] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in Portuguese", in *Proc. ICASSP*, 2008.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update", *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[22] J. Kim, S. E. Schwarm, and M. Ostendorf, "Detecting structural metadata with decision trees and transformation-based learning", in *HLT-NAACL*, 2004.