Disfluency Detection Based on Prosodic Features for University Lectures

Henrique Medeiros[1,2], *Helena Moniz*[1,3], *Fernando Batista*[1,2], *Isabel Trancoso*[1,4], *Luis Nunes*[2,5]

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

hrbmedeiros@hotmail.com, {helenam;fmmb;imt}@l2f.inesc-id.pt, luis.nunes@iscte.pt

Abstract

This paper focuses on the identification of disfluent sequences and their distinct structural regions, based on acoustic and prosodic features. Reported experiments are based on a corpus of university lectures in European Portuguese, with roughly 32h, and a relatively high percentage of disfluencies (7.6%). The set of features automatically extracted from the corpus proved to be discriminant of the regions contained in the production of a disfluency. Several machine learning methods have been applied, but the best results were achieved using Classification and Regression Trees (CART). The set of features which was most informative for cross-region identification encompasses word duration ratios, word confidence score, silent ratios, and pitch and energy slopes. Features such as the number of phones and syllables *per* word proved to be more useful for the identification of the interregnum, whereas energy slopes were most suited for identifying the interruption point.

Index Terms: prosodic features, automatic disfluency detection, corpus of university lectures, machine learning.

1. Introduction

Automatic speech recognition systems (ASR) have recently been conquering their place in the information society, and are now being applied for well-known tasks, like automatic subtitling, speech translation, speech summarization and production of multimedia content. However, speech is a rich source of information, from which a vast number of structural phenomena can be extracted. Enriching the ASR output with such structural phenomena is crucial for improving the human readability, for further automatic processing tasks, and also opens new horizons to possible application. Disfluencies characterize speech and play a special role as a structural phenomena. Considering them becomes indispensable in the development of a robust transcription system, because: i) they may trigger readability issues caused by an interruption of the normal flow of an intended message, ii) they provide crucial clues for characterizing the speaker and speaking styles, and iii) they are also relevant to disambiguate possible locations of sentence-like units in speech.

This paper focuses on the prediction of disfluent sequences in a corpus of university lectures in European Portuguese (EP) and on the characterization of the distinct disfluent regions contained in a disfluent sequence. The specific domain is very challenging, mainly because we are dealing with quite informal lectures, contrasting with other corpus already collected of more formal seminars. This is the first work conducted for EP that aims at predicting all categories of disfluent events based exclusively on the automatic audio-segmentation and prosodic features, using distinct classification methods to evaluate the best performance achieved. Moreover, it is also a step-forward in automatically characterizing all the regions of a disfluent sequence.

2. Related work

Disfluent sequences have a structure composed of several possible regions: a region to be auto-corrected, the reparandum; a moment where the speaker interrupts his/her production, known as the interruption point (IP); an optional editing phase or interregnum, filled with expressions such as *aa*/"uh" or *vocês sabem, percebem*/"you know"; and a repair region, where speech fluency is recovered [1, 2, 3]. Determining such structural elements is not a trivial task [3], but it is known that speakers signal different cues in those regions [4] and several studies have found combinations of cues that can be used to identify disfluencies and repairs with reasonable success [3, 5]. According to [3, 5, 6], based on the analysis of several disfluent types, those cues may relate to segment duration, intonation characteristics, word completion, voice quality alternations, and pattern coarticulations [6]. According to [7, 8] fragments can be problematic for recognition if not considered and fairly identified. In a different perspective they are also referred to as important cues to disfluent regions identifiable throughout prosodic features. Even thought fragments are common in human speech, [9] shows that they can present different significant characteristics across languages. Filled pauses are also problematic since they can be confused and recognized as small functional words, resulting in structures that decrease the ASR performance.

For European Portuguese, only recently a reduced number of studies on characterizing disfluencies have been conducted. [10] analyze the acoustic characteristics of filled pauses *vs.* segmental prolongations in a corpus of Portuguese broadcast news, using prosodic and spectral features to discriminate between both categories. [11, 12] use the same university lectures corpus subset also used in the present study and concluded that the best features to identify whether an element should be rated as fluent or disfluent are: prosodic phrasing, contour shape, and presence/absence of silent pauses. Recently, [13] analyze the prosodic behavior of the different regions of a disfluency se-

Copyright © 2013 ISCA

Corpus subset →	train+dev	test
Time (h)	28:00	3:24
number of disfluent sequences	8390	950
number of words + filled pauses	216435	24516
number of elements in a disfluency	16360	2043
elements in disfluencies (%)	7.6	8.3
filled pauses in disfluencies (%)	23.5	18.0
fragments in disfluencies (%)	10.9	11.3
disfluencies containing IP (%)	34.9	35.2
disfluencies with interregnum (%)	23.5	18.0
disfluencies followed by repair (%)	34.7	35.2

Table 1: Properties of the Lectra training subset.

quence, pointing out to prosodic contrast strategy (pitch and energy increases) between the reparandum and the repair. The authors evidenced that, although prosodic contrast marking between those regions is a cross speaker and cross category strategy, there are degrees in doing so, namely: filled pauses exhibit the highest f_0 increase, and repetitions exhibit the highest energy. Regarding temporal patterns, [14] show that the disfluency is the longest event, the silent pause between the disfluency and the following word is longer in average than the previous one, and that the first word of a repair equals the silent pause before a disfluency, being the shortest events.

Different methods have been proposed for the classification of disfluent regions, but the use of Classification and Regression Trees (CART) is usually considered to be a good choice [3, 15, 16]. In contrast to single model usage multi-method classifications as well as multi-knowledge sources usually result in better predictions [7, 17, 18, 19].

3. Corpus

This work is based on Lectra, a speech corpus of university lectures in European Portuguese, originally created for multimedia content production and to support hearing-impaired students [20]. The corpus contains records from seven 1-semester courses, where most of the classes are 60-90 minutes long, and consist of spontaneous speech mostly.Due to a recent extension, its current version contains about 32h of manual orthographic transcripts and was split into 2 different subsets (training+development and test) [21]. Overall statistics about this corpus are presented in Table 1.

Along with the manual transcripts we also have available force aligned and automatic transcripts, produced by the in-house ASR Audimus [22]. The ASR was trained for the Broadcast News domain and for that reason it presents a word error rate (WER) of about 50%. The high WER and the scarcity of text materials in our language to train language models for the university lectures domain has motivated the decision of using the ASR also in a forced alignment mode, in order not to bias the study with the poor results obtained with an out-of-domain recognizer. For sake of comparison, all the results will be reported for both force aligned and automatic transcripts. The corpus is available as self-contained XML files [23]. Each XML corresponds to a transcript integrating both manual and automatic synchronized transcripts, enriched with additional prosodic information related to pitch, energy, duration, and other structural metadata (punctuation, disfluencies, paralinguistic annotation, etc.).

4. Feature set

In order to use the XML files, a parser was created that allows not only to extract pre-stored information, but also to compute more complex features. The following features were used either for the current word (cw) or for the following word (fw): $conf_{cw}$, $conf_{fw}$ (ASR confidence scores), dur_{cw}, dur_{fw} (word durations), $phones_{cw}$, $phones_{fw}$ (number of phones), syl_{cw}, syl_{fw} (number of syllables), $pslope_{cw}$, $pslope_{fw}$ (pitch slopes), $eslope_{cw}$, $eslope_{fw}$ (energy slopes), [$pmax_{cw}$, $pmin_{cw}$, $pmed_{cw}$ (pitch maximum, minimum, and median)], $emax_{cw}$, $emin_{cw}$, $emed_{cw}$ (energy maximum, minimum and median), $bsil_{cw}$, $bsil_{fw}$ (silences before the word). The following features involving two consecutive words were calculated: $equals_{pw,cw}$, $equals_{cw,fw}$ (binary features indicating equal words), $sil.cmp_{cw,fw}$ (silence comparison), $dur.cmp_{cw,fw}$ (duration comparison), $pslopes_{cw,fw}$ (shape of the pitch slopes), $eslopes_{cw,fw}$ (shape of the energy slopes), $pdif_{pw,cw}$, $pdif_{cw,fw}$, $edif_{pw,cw}$, $edif_{cw,fw}$ (pitch and energy differences), $dur.ratio_{cw,fw}$ (words duration ratio), $bsil.ratio_{cw,fw}$ (ratio of silence before each word), $pmed.ratio_{cw,fw}$, $emed.ratio_{cw,fw}$ (ratios of pitch and energy medians). Features within square brackets were used only in preliminary tests, but their contribution was not substantial and therefore were not used in subsequent experiments for simplification. In fact, some of the information contained in those features may be already encoded by the remaining features, such as slopes, shapes, and differences.

Pitch slopes were calculated based on semitones rather than frequency values. Slopes in general were calculated using linear regression. Silence and duration comparisons assume 3 possible values, expanding to 3 binary features: $>$ (greater than), $=$ (equal), or $<$ (less than). The pitch and energy shapes expand to 9 binary features, assuming one of the following values $\{RR, R-, RF, -R, --, -F, FR, F-, FF\}$, where $F = Fall$, $- = stationary$, $R = Rise$, and the i^{th} letter corresponds to the word i. The ratios assume values between 0 and 1, indicating whether the second value is greater than the first.

None of the above mentioned features uses lexical information, except for the feature that compares two words between them. However, this could also be replaced by an acoustic feature, since comparing two segments of speech can be performed fairly well on the acoustic level.

Apart from the previous automatic features, some experiments use two additional features that indicate the presence of fragments (FRG) and filled pauses (FP). We are currently using the manual classifications of those categories, but we also aim at verifying the impact of our set of features in the automatic identification of those categories. It is important to notice that while the automatic identification of fragments is still an active research area [16, 8], the automatic identification of filled pauses in spontaneous speech currently achieves an acceptable performance [24, 25].

5. Experiments and Results

This section presents four main experiments concerning the automatic detection of disfluencies and their structural elements. The first experiment aims at automatically identifying which words belong to a disfluent sequence. The second experiment aims at automatically identifying the IP (Interruption Point) of a disfluency, supported by findings that suggest that the IP is the major key in identifying the disfluent region [18]. A third and a fourth experiment identify the interregnum and the repair. A fi-

Conditions	Prec.	Rec.	F	SER
Align with FP&FRG	91.2	36.9	52.5	66.7
Align without FP&FRG	66.3	20.3	31.0	90.0
ASR	71.2	13.7	23.0	91.8

Table 2: Predicting elements that belong to disfluent sequences.

nal experiment distinguishes between five different regions: IP, interregnum, any other position in a disfluency, repair, or words outside a disfluency.

The evaluation is performed using standard performance metrics: Precision, Recall, F-measure and SER (Slot Error Rate) [26], which corresponds to the NIST error rate, used in the NIST Rich Transcription evaluations. Only elements that we aim at identifying are considered as slots and used by these metrics. Hence, for example, for the task of detecting the interruption point, the SER is computed by dividing the number of IP errors (misses and false alarms) by the number of IPs in the reference. Experiments here described were conducted using Weka[1], a collection of open source of machine learning algorithms and a collection of tools for data pre-processing and visualization. All experiments use 80% of the data for training while the remaining 20% are used for evaluation. Different classification algorithms were tested, namely: Naive Bayes, Logistic Regression, Multilayer Perceptron, and CART. All reported results were achieved using CARTs, which consistently achieved the best performance.

5.1. Detecting disfluent sequences

This set of experiments aims at automatically identifying words that belong to a disfluent sequence. Table 2 summarizes the results achieved by CARTs, using the set of features described above. The first two rows refer to experiments based on forced alignments, either including manual information about fragments (FRG) and filled pauses (FP), or not, respectively. The last row refers to results achieved for automatic speech transcripts. The performance is measured in terms of (Prec)ision, (Rec)all, (F)-measure, and SER, where each slot corresponds to elements marked as belonging to disfluent sequences. It is known that the initial words of a disfluency may be in fact fluent, since there are no cues at the onset of a reparandum, which contributes to making this task even more difficult. Not knowing whether the current element is a fragment or a filled pause may have a strong impact in the results. This can be seen in the first two rows of the results, which correspond to a reduction of the number of correctly classified elements belonging to a disfluency from 754 to 414. These two features are consensually described in the literature as having a major impact in the identification of the different disfluent regions. For instance, [3] states that in telephone conversations fragments occur in 60% of the regions to repair, and are therefore a reliable cue to identify the end of a reparandum. In our corpus, the percentage of fragments is much lower (10.9%), but they do have an impact on the results. The percentage of filled pauses (22.9%) is the largest of all disfluency types. [3] reports 89% precision and 78% for a similar task of detecting disfluencies in telephone conversations using a subset of our features (except for whether the current word is accented), but results apply to a different corpus and to a different language. Our results concerning automatic transcripts are mostly affected by the lower achieved recall.

[1]Weka version 3-6-8. http://www.cs.waikato.ac.nz/ml/weka

Conditions	Prec.	Rec.	F	SER
Align with FP&FRG	77.6	36.7	49.8	73.9
Align without FP&FRG	71.4	11.8	20.3	92.9
ASR	73.9	2.4	4.7	98.4

Table 3: Predicting the Interruption Point.

Conditions	Prec.	Rec.	F	SER
Align with FP&FRG	96.8	99.7	98.3	3.5
Align without FP&FRG	69.2	42.2	52.5	76.6
ASR	76.8	80.6	78.7	43.7

Table 4: Predicting the Interregnum.

5.2. Detecting the Interruption Point

In our corpus, about 35% of the disfluent sequences account for the existence of an IP. Table 3 shows the performance achieved for task also using CARTs, revealing a significantly lower performance when comparing with the previous task. Results indicate that fragments and filled pauses are crucial for the identification of IPs, affecting specially the recall since most IPs are followed by filled pauses. This task is often reported in the literature as being performed in a multi-pass fashion, where a first pass corresponds to identifying fragments and filled pauses, and the second pass uses previous identification results as well as lexical matches between the reparandum and the repair for identifying the IP and for segmenting the reparandum and the repair properly [27, 18].

5.3. Detecting the *Interregnum*

This task aims at identifying which elements in a corpus match the interregnum of a disfluency, which roughly corresponds to finding the filled pauses that occur at the final positions of a disfluent sequence. In our corpus, interregnum accounts almost exclusively for filled pauses. Editing expressions (*quer dizer*/"I mean") correspond to only 12 cases in the whole corpus, and there is a strong tendency for the non co-occurence of discourse markers (*pronto, portanto*/"so", *portanto*/"like", *vocês sabem, percebem*/"you know", etc.) with filled pauses. Therefore, knowing the location of filled pauses yields a performance close to 100% for this task. Table 4 presents the results. When no information about filled pauses and fragments is given, the performance is strongly affected. Results for automatic transcripts are surprisingly good, despite the high WER, which suggests that cues about filled pauses can still be found in the data.

5.4. Detecting the repair

The repair is of particular interest, because it is often difficult to distinguish from a punctuation mark or from a sentence boundary. Table 5 shows the achieved results, revealing a poor recall performance, specially in the automatic transcripts where only 13 of the 720 possible repairs were correctly identified.

Conditions	Prec.	Rec.	F	SER
Align with FP&FRG	70.1	12.4	21.0	92.9
Align without FP&FRG	69.2	11.3	19.4	93.8
ASR	61.9	1.9	3.6	99.3

Table 5: Predicting the repair region.

Conditions	Prec.	Rec.	F	SER
Align with FP&FRG	81.5	27.6	41.2	75.0
Align without FP&FRG	58.6	15.0	23.9	90.7
ASR	64.6	9.9	17.2	94.0

Table 6: Predicting all the distinct elements.

Conditions	Prec.	Rec.	F	SER
inDisf	38.7	1.3	2.4	100.7
IP	53.4	15.1	23.6	98.1
interregnum	61.2	51.5	55.9	81.2
repair	64.2	14.4	23.6	93.6

Table 7: Individual performance results *per* task.

5.5. Distinguishing between all the structural elements

This final set of experiments consist of a multiclass detection that distinguishes between all the above-mentioned structural. Table 6 shows the corresponding overall results. Details on the performance achieved for each task are also presented in Table 7 and in the confusion matrix from Table 8, which consider forced alignments without information about filled pauses and fragments. *inDisf* corresponds to words inside the disfluency that do not match the IP or the interregnum. The performance for each of the individual structures is even better than performing each one separately, but the confusion matrix reveals that the results are still much influenced by the number of deletions. The overall performance is affected by the low detection performance for *inDisf* words, because most of such words are in fact fluent and thus difficult to distinguish from words outside of a disfluency [3, 6].

5.6. Feature Analysis

To conclude this study, we have analyzed the impact of each feature in each of the previously described tasks. Table 9 shows the 20 most informative features for forced alignment where filled pauses and fragments were not used as features. While features like previous and current words being equal (2), duration ratio (5), and word confidence score (6) have a strong impact for all the tasks, features like the number of syllables (1), the current and following words being equal (2), and the number of phones (4) do have more impact in specific tasks. Shape of the pitch and energy slopes, and silence and duration comparisons (12-27) turned out to be very informative features. The remaining features, not shown in the table, also have an impact on the results even thought they are not represented. Filled pauses and fragments become the most relevant features when included as features, which confirms our expectations [3]. In addition, we have observed that the order of the most informative features is not significantly affected when such information is provided.

Classified as →	inDisf	IP	int	repair	*Del.*
inDisf	**12**	30	13	12	889
IP	10	**109**	23	4	574
interregnum	1	13	**189**	10	154
repair	1	3	16	**104**	596
Insertions	7	49	68	32	

Table 8: Confusion matrix without filled pauses and fragments.

	Feature	inDisf	IP	int.	repair	All
1	syl_{cw}	▪▪▪		▪▪▪▪▪		▪▪▪▪▪
2	$dur.ratio_{cw,fw}$	▪▪▪▪▪	▪▪▪▪▪	▪▪▪▪▪	▪▪▪▪▪	▪▪▪▪▪
3	$equals_{cw,fw}$	▪▪▪▪▪	▪▪▪▪▪		▪▪▪▪▪	▪▪▪▪
4	$bsil.ratio_{cw,fw}$		▪▪▪▪▪	▪▪▪▪▪	▪▪▪▪▪	▪▪▪▪▪
5	$pmed.ratio_{cw,fw}$	▪▪▪▪	▪▪	▪▪▪▪▪	▪▪▪▪	▪▪▪▪▪
6	$conf_{cw}$	▪▪▪▪▪	▪▪▪▪	▪▪▪▪	▪▪▪▪	▪▪▪
7	$equals_{pw,cw}$	▪▪▪▪▪	▪▪▪▪	▪▪▪	▪▪▪▪▪	▪▪▪▪
8	$emed.ratio_{cw,fw}$	▪▪▪▪	▪▪▪	▪▪▪▪	▪▪	▪▪▪▪
9	$phones_{cw}$	▪▪▪▪	▪▪	▪▪▪▪		▪▪▪▪
10	$eslopes : RR_{cw,fw}$	▪▪▪	▪▪▪▪▪	▪	▪▪	▪▪
11	$sil.cmp :>_{cw,fw}$	▪▪▪	▪▪▪▪	▪▪▪▪	▪▪▪	▪▪▪
12	$conf_{fw}$		▪▪	▪	▪▪▪▪	▪▪▪▪
13	$sil.cmp :<_{cw,fw}$	▪▪▪▪	▪▪▪	▪▪▪▪	▪▪	▪▪▪
14	$eslopes : FF_{cw,fw}$	▪▪▪	▪▪▪▪	▪▪	▪▪▪▪	▪
15	$pslopes : R-_{cw,fw}$	▪▪▪	▪	▪▪▪	▪▪▪▪	▪▪▪
16	$sil.cmp :=_{cw,fw}$	▪▪▪▪	▪▪▪▪	▪▪▪	▪▪	▪▪
17	$pslopes : FR_{cw,fw}$	▪▪	▪▪	▪▪▪	▪	▪▪▪
18	$pslopes : RF_{cw,fw}$	▪	▪	▪▪	▪▪▪	▪▪
19	$eslopes : RF_{cw,fw}$	▪	▪▪▪	▪	▪	▪
20	$eslopes : R-_{cw,fw}$				▪▪▪	

Table 9: Top 20 most influent features, not considering fragments and filled pauses.

6. Conclusions

This paper presents a number of experiments focusing on the automatic identification of disfluent sequences, and on distinguishing between their structural elements. To the best of our knowledge this is the first work that automatically identifies disfluencies and their structural elements for a Portuguese corpus, and represents an important step in the development of this kind of systems for our language. The performance achieved for detecting words inside of disfluent sequences is about 91% Precision and 37% Recall, when filled pauses and fragments are used as a feature. Presented results confirm that knowledge about filled pauses and fragments has a strong impact on the performance. Without it, the performance decays to 66% Precision and 20% Recall. Results also suggest that the interregnum is the easiest structural element to identify, even when no filled pauses and fragments are used as features. That was also observed on automatic transcripts, created by an out-of-domain speech recognition. The proposed features were able to detect the IP and the repair at a reasonable precision (53% and 64%, respectively), but the overall performance is affected by a low recall. The relevance of individual features for each of the tasks has been analyzed, showing that word confidence scores, word duration ratio, and knowing when words are equal, have an overall strong impact, while features, such as number of syllables, and number of phones have more impact in specific tasks. Moreover, shapes of the pitch and energy slopes, and silence and duration comparisons proved to be very informative.

Future experiments will focus on performing similar experiments with two existent Portuguese corpora (broadcast news and map-task), complementing the on-going cross-domain analysis.

7. 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 project PEst-OE/EEI/LA0021/2013, by DIRHA European project FP7-ICT-2011-7-288121, and by ISCTE – IUL.

8. References

[1] W. Levelt, "Monitoring and self-repair in speech," *Cognition*, no. 14, pp. 41–104, 1983.

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

[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] D. Hindle, "Deterministic parsing of syntactic nonfluencies," in *ACL*, 1983, pp. 123–128.

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

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

[7] A. S. Yang Liu, Elizabeth Shriberg, "Automatic disfluency identification in conversational speech using multiple knowledge sources," in *EUROSPEECH 2003 - INTERSPEECH 2003*, 2003.

[8] J.-F. Yeh and M.-C. Yen, "Speech recognition with word fragment detection using prosody features for spontaneous speech," *Applied Mathematics and Information Sciences*, 2012.

[9] C.-T. Chu, Y.-H. Sung, Y. Zhao, and D. Jurafsky, "Detection of word fragments in mandarin telephone conversation," in *INTERSPEECH*, 2006.

[10] A. Veiga, S. Candeias, C. Lopes, and F. Perdigão, "Characterization of hesitations using acoustic models," in *International Congress of Phonetic Sciences - ICPhS XVII*, 2011.

[11] 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.

[12] H. Moniz, F. Batista, I. Trancoso, and A. I. Mata, *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, 1st ed., ser. Lecture Notes in Computer Science. Caserta, Italy: Springer Berlin / Heidelberg, January 2011, vol. 6456, ch. Analysis of interrogatives in different domains, pp. 136–148.

[13] H. Moniz, F. Batista, I. Trancoso, and A. I. M. da Silva, "Prosodic context-based analysis of disfluencies," in *In Interspeech 2012*, 2012.

[14] H. Moniz, F. Batista, A. Mata, and I. Trancoso, "Analysis of disfluencies in a corpus of university lectures," in *Proc. of Exling*, Athens, Greece, 2012.

[15] E. Shriberg, R. Bates, and A. Stolcke, "A prosody only decision tree model for disfluency detection," in *Proc. EUROSPEECH*, 1997.

[16] Y. Liu, "Word fragment identification using acoustic-prosodic features in conversational speech," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Proceedings of the HLT-NAACL 2003 student research workshop - Volume 3*, ser. NAACLstudent '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 37–42. [Online]. Available: http://dx.doi.org/10.3115/1073416.1073423

[17] D. Baron, E. Shriberg, and A. Stolcke, "Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues," in *in Proc. of the International Conference on Spoken Language Processing*, 2002, pp. 949–952.

[18] 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.

[19] M. Snover, B. Dorr, and R. Schwartz, "A lexically-driven algorithm for disfluency detection," in *Proceedings of HLT Conference/ NAACL annual meeting*, 2004.

[20] 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.

[21] T. Pellegrini, H. Moniz, F. Batista, I. Trancoso, and R. Astudillo, "Extension of the lectra corpus: classroom lecture transcriptions in european portuguese," in *SPEECH AND CORPORA*, 2012.

[22] 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.

[23] F. Batista, H. Moniz, I. Trancoso, N. Mamede, and A. Mata, "Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation," *Journal of Speech Sciences*, vol. 2, no. 2, pp. 115–138, November 2012. [Online]. Available: http://www.journalofspeechsciences.org/index.php/journalofspeechsciences/article/view/60

[24] D. O'Shaughnessy, "Recognition of hesitations in spontaneous speech," in *Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1*, ser. ICASSP'92. Washington, DC, USA: IEEE Computer Society, 1992, pp. 521–524. [Online]. Available: http://dl.acm.org/citation.cfm?id=1895550.1895694

[25] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *In Proceedings of Eurospeech '99*, 1999, pp. 227–230.

[26] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. of the DARPA Broadcast News Workshop*, Herndon, VA, Feb. 1999.

[27] J. Kim, S. E. Schwarm, and M. Ostendorf, "Detecting structural metadata with decision trees and transformation-based learning," in *Proc. HLT-NAACL*, 2004, pp. 137–144.