Comparing Different Machine Learning Approaches for Disfluency Structure Detection in a Corpus of University Lectures*

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

1 **Laboratório de Sistemas de Língua Falada - INESC-ID, Lisboa, Portugal**
ISCTE - Instituto Universitário de Lisboa, Lisboa, Portugal
hrbmedeiros@hotmail.com, Fernando.Batista@iscte.pt

2 **Laboratório de Sistemas de Língua Falada - INESC-ID, Lisboa, Portugal**
FLUL/CLUL, Universidade de Lisboa, Lisboa, Portugal
helena.moniz@inesc-id.pt

3 **Laboratório de Sistemas de Língua Falada - INESC-ID, Lisboa, Portugal**
Instituto Superior Técnico (IST), Lisboa, Portugal
isabel.trancoso@inesc-id.pt

4 **ISCTE - Instituto Universitário de Lisboa, Lisboa, Portugal**
Instituto de Telecomunicações, Lisboa, Portugal
luis.nunes@iscte.pt

Abstract

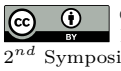
This paper presents a number of experiments focusing on assessing the performance of different machine learning methods on the identification of disfluencies and their distinct structural regions over speech data. Several machine learning methods have been applied, namely Naive Bayes, Logistic Regression, Classification and Regression Trees (CARTs), J48 and Multilayer Perceptron. Our experiments show that CARTs outperform the other methods on the identification of the distinct structural disfluent regions. Reported experiments are based on audio segmentation and prosodic features, calculated from a corpus of university lectures in European Portuguese, containing about 32h of speech and about 7.7% of disfluencies. The set of features automatically extracted from the forced alignment corpus proved to be discriminant of the regions contained in the production of a disfluency. This work shows that using fully automatic prosodic features, disfluency structural regions can be reliably identified using CARTs, where the best results achieved correspond to 81.5% precision, 27.6% recall, and 41.2% F-measure. The best results concern the detection of the interregnum, followed by the detection of the interruption point.

1998 ACM Subject Classification I.2.7 Natural Language Processing

Keywords and phrases Machine learning, speech processing, prosodic features, automatic detection of disfluencies

Digital Object Identifier 10.4230/OASIcs.SLATE.2013.259

* This work was supported by national funds through FCT – Fundação para a Ciência e a Tecnologia – under Ph.D grant SFRH/BD/44671/2008, parcially supported by projects CMU-PT/HuMach/0039/2008 and PEst-OE/EEI/LA0021/2011, and also by DCTI - ISCTE – Instituto Universitário de Lisboa.



2^{nd} Symposium on Languages, Applications and Technologies (SLATE'13).
Editors: José Paulo Leal, Ricardo Rocha, Alberto Simões; pp. 259–269

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Disfluencies are a linguistic mechanism used for on-line editing a message. Disfluencies encompass several distinct types, namely, filled pauses, prolongations, repetitions, deletions, substitutions, fragments, editing expressions, insertions or complex sequences (more than one category uttered) [27]. Those events have been studied from different perspectives, in Psycholinguistics, in Linguistics, in Text-to-speech, and in Automatic Speech Recognition (ASR). The latter will be the focus of our study, since it is well-known that disfluencies are a challenging structure for ASR systems, mainly due to the fact that they are not well recognized and the adjacent words are also influenced and may be erroneously identified.

Automatic speech recognition systems have recently earned 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. Speech is a rich source of information from which a vast number of structural phenomena can be extracted, apart from a text stream. Enriching the ASR output with structural phenomena is crucial for improving the human readability, for further automatic processing tasks, and also opens new horizons to a vast range of applications. Disfluencies characterize spontaneous and prepared speech and play a special role as a structural phenomena in speech [12, 4, 6]. Considering them becomes indispensable in the development of a robust and natural ASR systems, 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, the speaking styles and iii) also in combination with segmentation tasks, they provide better sentence-like units detection.

This paper analyses the performance of different machine learning methods on the prediction of disfluent sequences and their distinct regions in a corpus of university lectures in European Portuguese. This paper complements the analysis performed in the scope of the work described in [17], where, for the first time, results for disfluency detection on Portuguese university lectures were presented. The specific domain is very challenging, mainly due to the fact that it comprehends quite informal lectures, contrasting with other data already collected of more formal seminars [10].

The chosen algorithms represent state-of-the-art machine learning techniques and are widely used by the scientific community for similar problems. The choice of methods was limited to a subset of methods available in the Weka suite, but other methods currently not available could also be explored, including CRFs (Conditional Random Fields), a promising method for sequence modeling. CARTs, in particular, have been widely adopted for related tasks in the literature [33, 30, 1, 32, 13, 22]. The purpose of this study is to assess the performance of the different methods, and reveal their strengths and weaknesses on the task of identifying the regions of a disfluency.

This paper is organized as follows: Section 2 overviews the literature concerning the detection of disfluencies and corresponding methods. Section 3 describes the *corpus* used in our experiments as well as the multilayer information available. Section 4 describes the adopted features. Section 5 describes the performance metrics that have been used for the evaluation. Section 6 presents experiments for either detecting elements that belong to a disfluent sequence, or distinguishing between those elements. Section 7 points out the major conclusions and presents issues still open for future work.

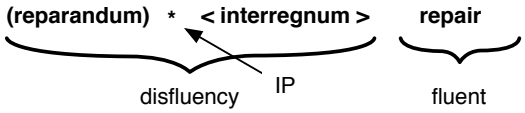


Figure 1 Different regions related to a disfluent sequence.

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 "uh" or "you know"; and a *repair* region, where speech fluency is recovered [11, 27, 22]. Figure 1 illustrates such structure. Determining such structural elements is not a trivial task [22, 34], but it is known that speakers signal different cues in those regions [9] and several studies have found combinations of cues that can be used to identify disfluencies and repairs with reasonable success [22, 7]. According to [29, 22, 7], based on the analysis of several disfluent types, those cues may relate to segment duration, intonation characteristics, word completion, voice quality alternations, vowel quality and co-articulation patterns [29]. According to [13, 38] 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 [38]. Even thought fragments are common in human speech, [3] shows that they can present different significant characteristics across languages. Filled pauses are also problematic since they can be confused and recognized as functional words, usually resulting in fragment-like structures that decrease the ASR performance [5, 28]. The potential benefit of modeling disfluencies in a speech recognizer in Spanish has been studied by [26], following a data driven approach.

For European Portuguese, only a recent and a reduced number of studies on characterizing disfluencies have been found in the literature. [36] 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. Slight pitch descendent patterns and temporal characteristics are pointed out as the best cues for detecting these two categories. [21, 20] use the same university lectures corpus subset also used in the present study and concluded that the best features to identify if a disfluency should be rated as either a fluent or a disfluent are: prosodic phrasing, contour shape, and presence/absence of silent pauses. Recently, [19] analyze the prosodic behavior of the different regions of a disfluency sequence, 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, meaning, filled pauses exhibit the highest f_0 increase and repetitions the highest energy one. Regarding temporal patterns, [18] 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 similar tasks in the literature, either generative or discriminative. The scientific community often assumes the CARTs produce good results, therefore being the preferred choice. In contrast to single model usage multi-method classifications as well as multi-knowledge sources usually result in better predictions [13, 1, 15, 31, 37].

Table 1 Properties of the Lectra training subset.

Corpus subset →	train+dev	test
Time (h)	28:00	3:24
Number of sentences	8291	861
Number of disfluencies	8390	950
Number of words (including filled pauses and fragments)	216435	24516
Number of elements inside a disfluency	16360	2043
Percentage of elements inside disfluencies	7.6%	8.3%

3 Data

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 [35]. The corpus contains records from seven 1-semester courses, where most of the classes are 60-90 minutes long, and consist mostly of spontaneous speech. It has been recently extended, now containing about 32h of manual orthographic transcripts [25]. Experiments here described use approximately 28h of the corpus to train models, and the remaining portion for testing. Table 1 presents overall statistics about the data.

Besides the manual transcripts, we also have available force-aligned transcripts, automatically produced by the in-house ASR Audimus [23]. The ASR used in this study was trained for the Broadcast News domain, therefore unsuitable for the university lectures domain. The scarcity of text materials in our language to train language models for this domain has motivated the decision of using the ASR in a forced alignment mode, in order not to bias the study with the poor results obtained with an out-of-domain recognizer. The corpus is available as self-contained XML files [2] that includes not only all the information provided by the speech recognition, but also the manually annotated information like punctuation marks, disfluencies, inspirations, etc. Each XML file also includes information related to pitch, energy, duration that comes from the speech signal and that has been assigned to different units of analysis, such as words, syllables and phones.

4 Feature Set

An XML parser was specially created with the purpose of extracting and calculating features from the XML files described in the previous section. The following features were extracted 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), [$pmax_{cw}$, $pmin_{cw}$, $pmed_{cw}$, $emed_{cw}$ (pitch maximum, minimum, and median; energy median)], $emax_{cw}$, $emin_{cw}$ (energy maximum and minimum), $bsil_{cw}$, $bsil_{fw}$ (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 expressed in brackets were used only in preliminary tests, but their contribution was not substantial and therefore, for simplification, they were not used in subsequent

experiments. It is important to notice that some of the information contained in the features that were not used in subsequent experiments is already encoded by the remaining features, such as slopes, shapes, and differences.

Pitch slopes were calculated based on semitones rather than raw 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. All the above features are based on audio segmentation and prosodic features, except for the feature that compares two consecutive words at the lexical level. In future experiments, we plan to replace it by an acoustic-based feature that compares two segments of speech on the acoustic level.

Apart from the previous automatic features, 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 [13, 38], the automatic identification of filled pauses in spontaneous speech has been applied with an acceptable performance [24, 8].

5 Evaluation Metrics

The following widely used performance evaluation metrics will be applied along the paper: Precision, Recall, F-measure, Slot Error Rate (SER) [16]. All these metrics are based on slots, which correspond to the elements that we aim at classifying. For example, for the task of classifying words as being part of a disfluency, a slot corresponds to a word marked as being part of a disfluency. Most of the results presented in the scope of this paper include all the standard metrics. However, F-measure is a way of having a single value for measuring the precision and the recall simultaneously and, as reported by [16], "this measure implicitly discounts the overall error rate, making the systems look like they are much better than they really are". For that reason, the preferred performance metric for performance evaluation will be the SER, which also corresponds to the NIST error rate used in their RT (Rich Transcription) evaluation campaigns. Notice, however, that SER is an error metric that assume values greater than 100 whenever the number of errors are greater than the number of slots in the reference.

The Receiver Operating Characteristic (ROC) is another performance metric, based on performance curves, that can also be used for more adequate analysis [14]. It consists of plotting the false alarm rate on the horizontal axis, while the correct detection rate is plotted on vertical. Most experiments reported in this paper also include a ROC value that corresponds to the area under the ROC curve.

6 Experiments and Results

Experiments here described were conducted using Weka[1], a collection of open source machine learning algorithms and a collection of tools for data pre-processing and visualization. Different classification algorithms were tested, namely: Naive Bayes, Logistic Regression, Multilayer Perceptron, CARTs and J48. For each one of the tested algorithms, the default parameters where were used.

The remainder of this section presents two complementary studies concerning the automatic detection of disfluencies and the identification of their structural elements, where the focus lies on comparing the results achieved with different methods. The first study involves a binary classification and aims at automatically identifying which words belong to a disfluent sequence. The second study comprises a multiclass classification that aims at distinguishing between five different regions related with disfluencies: IP, interregnum, any other position in a disfluency, repair, any other position outside a disfluency. Concerning the multiclass classification, details relative to distinct disfluent zone classification performance will be presented.

6.1 Detecting Elements belonging to Disfluent Sequences

This first set of experiments aims at automatically identifying words that belong to a disfluency. Table 2 summarizes the overall performance results, in terms of time taken and correctly classified instances, for binary predicting whether a word (including filled pauses and fragments) belongs to a disfluent sequence or not. Each column represents results for a distinct algorithm, namely: baseline achieved by simply selecting the most common prediction (ZeroR), Naive Bayes (NB), Logistic Regression (LR), Classification and Regression Tree (CART), MultiLayer Perceptron (MLP) and J48. The percentage of Correctly Classified Instances takes into account all the elements that are being classified, and not only slots (*vide* Section 5). The baseline achieved using ZeroR (91.7%) corresponds to marking all words as being outside of a disfluency, which is consistent with the percentage of elements in the test corpus belong to disfluencies (*vide* Table 1). The value referred as Kappa indicates whether a classifier is doing better than chance. The last two lines of the table reveal that both Logistic Regression and CARTs are the most promising approaches. The time taken to build the model is considerable less for Logistic Regression, when compared with the other methods. In fact, Logistic Regression is approximately 85 times faster when compared to CART, and the other performance results presented in the table are quite similar.

The detailed performance results for each method based on slots are also presented in Table 3, where each slot corresponds to elements marked as being part of a disfluency. The first 3 columns report the actual counts for *Correct*, *Inserted* (not marked in the reference),

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

Table 2 High level performance analysis for predicting words that belong to disfluencies.

	ZeroR	NB	LR	CART	MLP	J48
Time taken to build the model (seconds)	0.1	551.9	40.5	3412.4	8473.2	3818.5
Time taken to test the model (seconds)	3.2	5.6	3.1	2.0	9.2	1.9
Correctly classified instances (%)	91.7	89.8	94.4	94.4	93.9	94.4
Kappa	0.0	0.362	0.503	0.502	0.489	0.505

Table 3 Detailed performance analysis on predicting words that belong to disfluencies.

Method	Cor	Ins	Del	Precision	Recall	F	SER	ROC
Naive Bayes	891	1339	1152	40.0	43.6	41.7	121.9	0.771
Logistic Regression	765	95	1278	89.0	37.4	52.7	67.2	0.797
CART	754	73	1289	91.2	36.9	52.5	66.7	0.726
MultiLayer Perceptron	799	244	1244	76.6	39.1	51.8	72.8	0.779
J48	778	115	1265	87.1	38.1	53.0	67.5	0.733

Table 4 High level performance analysis for a multiclass prediction.

	ZeroR	NB	LR	CART	MLP	J48
Time taken to build the model (secs)	0.1	574.7	1391.1	6148.8	10209.7	4602.1
Time taken to test the model (secs)	3.4	7.4	3.9	1.8	12.3	1.9
Correctly classified instances (%)	88.7	76.5	91.4	91.5	91.4	91.4
Kappa	0.0	0.223	0.416	0.420	0.414	0.414

and *Deleted* (marked in the reference but not correctly classified) slots. Values presented for *Precision*, *Recall*, *F-measure* and SER (*vide* Section 5) represent percentages. Because CARTs are not probabilistic classifiers, the ROC value can not be fairly computed, and for that reason it was not presented. Results reveal that CART and Logistic Regression present the best performance values, where CARTs achieved a better precision and Logistic Regression achieved a better recall. It is interesting to notice that while the F-measure is better for the Logistic Regression, the SER is the best for CART, which might be a more meaningful measure.

6.2 Distinguishing between all the Structural Elements

This set of experiments aims at identifying the structural elements that compose or are related to a disfluency. Table 4 summarizes the overall performance results, in terms of time taken and correctly classified instances. The time taken to build the model is considerable less for Naive Bayes, but the performance is above the baseline achieved using ZeroR. Performing Logistic Regression is also less time consuming than the other three methods, but such difference is now less notorious than before. The values presented in the last two rows suggest that all approaches (except Naive Bayes) achieve similar performances, and that CARTs achieve the best results by a small difference.

Table 5 presents a more detailed analysis of the performance of each one of the approaches, revealing that CART should be the best choice for this type of problem. The table also includes the number of substitutions (Sub), which correspond to the number of mistakes

Table 5 Detailed performance analysis for a multiclass prediction.

Method	Cor	Ins	Del	Sub	Precision	Recall	F-measure	SER
Naive Bayes	980	3983	1317	466	18.1	35.5	23.9	208.7
Logistic Regression	763	118	1883	117	76.5	27.6	40.6	76.7
CART	762	71	1899	102	81.5	27.6	41.2	75.0
MultiLayer Perceptron	753	99	1891	119	77.5	27.3	40.3	76.3
J48	749	96	1891	123	77.4	27.1	40.2	76.4

Table 6 Zone discrimination CART results.

	Cor	Ins	Del	Sub	Prec.	Recall	F	SER
IP	271	82	449	0	76.8	37.6	50.5	73.8
interregnum	366	12	1	0	96.8	99.7	98.3	3.5
other word inside disfluency	19	33	937	0	36.5	2.0	3.8	101.5
repair	106	46	614	0	69.7	14.7	24.3	91.7
outside disfluency	*21682*	*23581*	*43435*	*0*	*47.9*	*33.3*	*39.3*	*102.9*
Overall performance	762	71	1899	102	81.5	27.6	41.2	75.0

Table 7 Cart confusion matrix.

Classified as →	IP	interregnum	in-disf	repair	outside disf
IP	271	0	19	5	425
interregnum	0	366	0	0	1
other word inside disfluency	58	0	19	14	865
repair	0	3	3	106	608
outside disfluency	24	9	11	27	*21682*

between the different possible slots. The best precision is by far achieved using a CART and Logistic Regression achieved the second best performance, and all metrics reflect this difference coherently.

6.2.1 Detailed CART Results

Taking into account that the best results previously presented concern CARTs, this section presents detailed performance results obtained with this approach. The best results for automatically identifying each one of the structural elements that are related with disfluencies are detailed in Table 6. The table reveals that, from all the structural elements related with a disfluency, the interregnum is by far the easiest to detect. That is an expected result because that information about filled pauses and fragments is being provided as a feature. All the presented results reveal a good precision when compared to recall except for interregnum. Good results considering both the F-measure and SER are also achieved for the detection of the IP. That is also not surprising, because the interruption point is often followed by filled pauses and sometimes preceded by fragments, for which our feature set includes information. The IP region is often referred as containing good clues for detecting disfluencies because the surrounding regions present characteristic contrasts in terms of feature values. Detecting the repair zone can also be performed at a considerably high precision, contrasting with the corresponding recall. A more deep word context analysis is needed to improve the recall performance on this classification. The worst classification refers to words that are marked as being part of a disfluent sequence, but not being neither the IP nor the interregnum, which correspond to words that most of the times are similar to fluent words. The line concerning the elements *outside a disfluency* refers to elements that were not considered one of the five possible structural elements of a disfluency, and correspond to non-slots.

The previous analysis can be complemented by also taking into consideration the corresponding confusion matrix, which is presented in Table 7. The matrix reveals that most of the elements are classified as being "outside of a disfluency", the most common situation in the corpus.

7 Conclusions

Different machine learning methods have been tested on the prediction of disfluent sequences and their distinct regions in a corpus of university lectures in European Portuguese. In terms of computational effort, Logistic Regression is the best choice, being much faster than the other classification approaches for binary predictions. Our experiments on the automatic identification of disfluent sequences suggest that similar results can be achieved using either CARTs or Logistic Regression. While CARTs tend to favor a better precision, Logistic Regression result in a better recall. Our experiments that distinguish between structural elements in a disfluent sequence suggest that CARTs are consistently better than the other tested approaches.

This paper complements the first studies that have been performed on detecting disfluencies and disfluency related regions for Portuguese university lectures [17]. For the future, we are planning a similar work for distinguishing between disfluency locations and punctuation marks.

References

1 Don Baron, Elizabeth Shriberg, and Andreas 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*, pages 949–952, 2002.

2 Fernando Batista, Helena Moniz, Isabel Trancoso, Nuno Mamede, and Ana Mata. Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. *Journal of Speech Sciences*, 2(2):115–138, November 2012.

3 Cheng-Tao Chu, Yun-Hsuan Sung, Yuan Zhao, and Daniel Jurafsky. Detection of word fragments in mandarin telephone conversation. In *INTERSPEECH 2006*. ISCA, 2006.

4 H. Clark. *Using language*. Cambridge: Cambridge University Press, 1996.

5 Martin Corley and Oliver W. Stewart. Hesitation disfluencies in spontaneous speech: The meaning of um. *Language and Linguistics Compass*, 2(4):589–602, 2008.

6 R. Dufour, V. Jousse, Y. Estève, F. Béchet, and G. Linarès. Spontaneous speech characterization and detection in large audio database. In *13-th International Conference on Speech and Computer (SPECOM 2009)*, St Petersburg (Russia), 21-25 june 2009.

7 E. Shriberg. Phonetic consequences of speech disfluency. In *International Congress of Phonetic Sciences*, pages 612–622, 1999.

8 Masataka Goto, Katunobu Itou, and Satoru Hayamizu. A real-time filled pause detection system for spontaneous speech recognition. In *In Proceedings of Eurospeech '99*, pages 227–230, 1999.

9 D. Hindle. Deterministic parsing of syntactic non-fluencies. In *ACL*, pages 123–128, 1983.

10 L. Lamel, G. Adda, E. Bilinski, and J. Gauvain. Transcribing lectures and seminars. In *Interspeech 2005*, Lisbon, Portugal, September 2005.

11 W. Levelt. Monitoring and self-repair in speech. *Cognition*, (14):41–104, 1983.

12 W. Levelt. *Speaking*. MIT Press, Cambridge, Massachusetts, 1989.

13 Yang 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*, NAACLstudent '03, pages 37–42, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

14 Yang Liu and Elizabeth Shriberg. Comparing evaluation metrics for sentence boundary detection. In *Proc. of the IEEE ICASSP*, Honolulu, Hawaii, 2007.

15 Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1526–1540, 2006.

16 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.

17 Henrique Medeiros, Helena Moniz, Fernando Batista, Isabel Trancoso, and Luis Nunes. Disfluency detection based on prosodic features for university lectures. Interspeech 2013 (submitted).

18 Helena Moniz, Fernando Batista, Ana Isabel Mata, and Isabel Trancoso. Analysis of disfluencies in a corpus of university lectures. In *ExLing 2012*, August 2012.

19 Helena Moniz, Fernando Batista, Isabel Trancoso, and Ana Isabel Mata da Silva. Prosodic context-based analysis of disfluencies. In *In Interspeech 2012*, 2012.

20 Helena Moniz, Fernando Batista, Isabel Trancoso, and Ana Isabel Mata. *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*, volume 6456 of *Lecture Notes in Computer Science*, chapter Analysis of interrogatives in different domains, pages 136–148. Springer Berlin / Heidelberg, Caserta, Italy, 1st edition edition, January 2011.

21 Helena Moniz, Isabel Trancoso, and Ana Isabel Mata. Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. In *Interspeech 2009*, Brighton, England, 2009.

22 C. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America (JASA)*, (95):1603–1616, 1994.

23 J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro. Broadcast news subtitling system in portuguese. In *ICASSP 2008*, pages 1561–1564, 2008.

24 Douglas O'Shaughnessy. Recognition of hesitations in spontaneous speech. In *Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1*, ICASSP'92, pages 521–524, Washington, DC, USA, 1992. IEEE Computer Society.

25 Thomas Pellegrini, Helena Moniz, Fernando Batista, Isabel Trancoso, and Ramon Astudillo. Extension of the lectra corpus: classroom lecture transcriptions in european portuguese. In *SPEECH AND CORPORA*, 2012.

26 Luis Javier Rodriguez Fuentes and M.I. Torres. Spontaneous speech events in two speech databases of human-computer and human-human dialogues in spanish. *Language and Speech*, 49(3):333–366, September 2006.

27 E. Shriberg. *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California, 1994.

28 Elizabeth Shriberg. Disfluencies in switchboard, 1996.

29 Elizabeth Shriberg. To "errrr" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31:153–169, 2001.

30 Elizabeth Shriberg, Rebecca Bates, and Andreas Stolcke. A prosody-only decision-tree model for disfluency detection. In *Proc. EUROSPEECH*, pages 2383–2386, 1997.

31 Matthew Snover and Bonnie Dorr. A lexically-driven algorithm for disfluency detection. In *in Proc. of HLT/NAACL*, pages 157–160, 2004.

32 Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Comput. Linguist.*, 26(3):339–373, September 2000.

33 Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Mari Ostendorf, Dilek Hakkani, Madelaine Plauche, Gokhan Tur, and Yu Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words, 1998.

34 Frederik Stouten, Jacques Duchateau, Jean-Pierre Martens, and Patrick Wambacq. Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation. *Speech Communication*, 48(11):1590–1606, 2006.

35 Isabel Trancoso, Rui Martins, Helena Moniz, Ana Isabel Mata, and Céu Viana. The lectra corpus - classroom lecture transcriptions in european portuguese. In *LREC*, 2008.

36 A. Veiga, S. Candeias, C. Lopes, and F. Perdigão. Characterization of hesitations using acoustic models. In *International Congress of Phonetic Sciences - ICPhS XVII*, volume -, pages 2054–2057, August 2011.

37 Andreas Stolcke Yang Liu, Elizabeth Shriberg. Automatic disfluency identification in conversational speech using multiple knowledge sources. *Trans. Audio, Speech and Lang. Proc.*, 17(7):1263–1278, September 2003.

38 Jui-Feng Yeh and Ming-Chi Yen. Speech recognition with word fragment detection using prosody features for spontaneous speech. *Applied Mathematics and Information Sciences*, 2012.