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 the performance of different machine learning methods on the identification of disfluencies and their distinct structural regions over speech data. Reported experiments are based on audio segmentation and prosodic features calculated from a corpus of university lectures in European Portuguese, containing about 24h of speech and about 7.5% 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. Several machine learning methods have been applied, namely Naive Bayes, Logistic Regression, Classification and Regression Trees (CARTs), and Multilayer Perceptron. Since the aim of the task is to perform a discriminative identification of the structural disfluent regions, CARTs outperform the others methods due to the very informed selection of the main features for each region. This work shows that using fully automatic prosodic features and CARTs disfluency structural regions can be reliably/suitably identified. The best results achieved using CARTs correspond to 83.6% precision, 32.5% recall, and 46.8 F-measure. All structural regions are being identified, but 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 – please refer to `http://www.acm.org/about/class/ccs98-html`

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

Digital Object Identifier 10.4230/OASIcs.xxx.yyy.p

* 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.

Conference/workshop/symposium title on which this volume is based on.
Editors: Billy Editor, Bill Editors; pp. 1–10

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

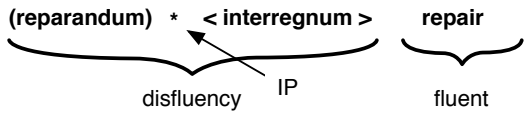

Figure 1 Different regions related to a disfluent sequence.

1 Introduction

Disfluencies are linguistic mechanism used to on-line editing of 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) [22]. Those events have been studied from different perspectives, in Psycholinguistics, in Linguistics, in Text-to-speech, 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 (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 play a special role as a structural phenomena in speech [7, 3]. 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[13], where for the first this results for portuguese university lectures were presented. The specific domain is very challenging, mainly due to the fact that we are dealing with quite informal lectures, contrasting with other data already collected of more formal seminars.

This paper is organized as follows: Section 2 overviews the literature concerning the detection of disfluencies and corresponding methods. Section 3 describes the data used in our experiments. Section 4 describes the features used. Section 5 describes the performance metrics that have been adopted for the evaluation. Section 6 presents details concerning each one of the experiments. Section 7 points out the major conclusions and presents the future work.

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 [6, 22, 18]. Figure 1 illustrate such structure. Determining such structural elements is not a trivial task [18], but it is known that speakers signal different cues in those regions [5] and several studies have found combinations of cues that can be used to identify disfluencies and repairs with reasonable success [18, 23]. According to [18, 23, 24], 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 pattern coarticulations [24]. According to [30, 31] 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, [2] shows that they can present different significant characteristics across languages. Filled pauses are also problematic since they can be confused and recognized as small words, usually resulting in fragment-like structures that decrease the ASR performance.

For European Portuguese, only a recent and a reduced number of studies on characterizing disfluencies have been found in the literature. [29] 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. [17, 16] 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, [15] 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, [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 similar tasks in the literature, either generative or discriminative. The scientific community often assumes the CARTs produce good results, therefore being the preferred choice [18, 25, 8]. In contrast to single model usage multi-method classifications as well as multi-knowledge sources usually result in better predictions [9, 27, 11, 26].

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 [28]. 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. It has been recently extended, now containing about 32h of manual orthographic transcripts [21]. Experiments here described use about 24h of the corpus, corresponding to about 78% of the whole corpus. Table 1 presents overall statistics about this subset.

Besides the manual transcripts we also have available force-aligned transcripts, automatically produced by the in-house ASR Audimus [19]. 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

Time (h)	24:28
Number of sentences	10576
Number of disfluencies	7382
Number of words (including filled pauses and fragments)	191210
Number of elements inside a disfluency	14357
Percentage of elements inside disfluencies	7.5%

Table 1 Properties of the Lectra training subset.

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 [1] that includes not only all the information provided by the speech recognition, but also the manually annotated information like punctuation marks, disfluencies, inspirations, etc. Information related to pitch, energy, duration, and that is enriched in terms of structural metadata. Each XML 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

All features were extracted or calculated from the above mentioned XML files by means of a parser, specially created for this purpose. The following pre-calculated features were produced 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 word 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 were not used in subsequent experiments for simplification. 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. 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, 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 [9, 32], the automatic identification of filled pauses in spontaneous speech has been performed with an acceptable performance [20, 4].

5 Evaluation Metrics

The following widely used performance evaluation metrics will be applied along the paper: Precision, Recall, F-measure, Slot Error Rate (SER) [12]. 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 correspond to a word marked as being part of a disfluency. Most of the results presented in the scope of this paper include the standard metrics: Precision, Recall, and F-measure. However, F-measure is a way of having a single value for measuring the precision and the recall simultaneously and, as reported by [12], "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 as an error metric assumes values greater than 100% whenever the number of errors are greater than the number of slots in the reference.

The ROC (Receiver Operating Characteristic) is another performance metric, based on performance curves, that can also be used for more adequate analysis [10]. 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. All experiences use 80% of the data for training while the remaining 20% are used for evaluation. For each tested algorithm initial parameters where left untouched. Different classification algorithms were tested, namely: Naive Bayes, Logistic Regression, Multilayer Perceptron, and CARTs.

The remainder of this section presents two experiments concerning the automatic detection of disfluencies and their structural elements, where the focus lies on comparing the results achieved with different methods. The first experiment describes a binary classification experiment that aims at automatically identifying which words belong to a disfluent sequence. The second experiment consists of 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. The best results, achieved using CARTs, will also be presented in detail.

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

	ZeroR	NB	LR	CART	MLP
Time taken to build the model (seconds)	0.1	211.2	33.8	1813.4	3364.8
Time taken to test the model (seconds)	0.3	2.0	0.5	0.2	3.3
Correctly classified instances (percentage)	92.5	90.7	95.4	95.5	94.8
Kappa	0.000	0.383	0.563	0.562	0.544

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

Method	Cor	Del	Ins	Precision	Recall	F	SER	ROC
Naive Bayes	1374	2090	1499	39.7	47.8	43.4	124.9	0.78
Logistic Regression	1234	118	1639	91.3	43.0	58.4	61.2	0.83
CART	1212	78	1661	94.0	42.2	58.2	60.5	-
MultiLayer Perceptron	1333	467	1540	74.1	46.4	57.1	69.9	0.80

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

6.1 Detecting elements belonging to disfluent sequences

This set of experiments aim at automatically identifying words that belong to a disfluency. Table 2 summarizes the overall results achieved for binary predicting whether a word belongs or not to a disfluent sequence. Each column represents results for a distinct algorithm, namely: simply selecting the most common prediction (ZeroR), Naive Bayes (NB), Logistic Regression (LR), CART, and MultiLayer Perceptron (MLP). The percentage of Correctly Classified Instances takes into account all the elements that are being classified. The baseline achieved using ZeroR (92.5%) corresponds to marking all words as being outside of a disfluency, since only 7.5% of all the elements in the corpus belong to disfluencies (*vide* Table 1). The value referred as Kappa indicates whether a classifier is doing better than chance. The 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 100 times faster when compared to MultiLayer Perceptron.

Table 3 presents performance details for each method based on slots, where each slot corresponds to words marked as being part of a disfluency. The first 3 columns report the actual counts for Correct, Deleted (marked in the reference but not correctly classified), and Inserts slots (not marked in the reference). Values presented for Precision, Recall, F-measure and Slot Error Rate represent percentages. Because CARTs are not probabilistic classifiers, the ROC value can not be fairly computed, and for that reason it was not presented. CART and Logistic Regression present the best performance values, and while CART achieved a better precision, Logistic Regression achieved a better recall. It is interesting to notice that while the F-measure is better for the Logistic Regression, the SER assumes the best value for the CART.

6.2 Distinguishing between all the structural elements

This set of experiments aim at identifying 5 structural elements related to disfluencies, and Table 4 summarizes the overall results. The time taken to build the model is considerable less for Naive Bayes, but also for Logistic Regression, when compared with the other two methods. However, such difference is now less notorious than before. The most promising approaches seem to be CARTs, Logistic Regression, and Multilayer Perceptron, based on the

	ZeroR	NB	LR	CART	MLP
Time taken to build the model (seconds)	0.1	326.5	636.3	4362.0	4267.3
Time taken to test the model (seconds)	0.2	6.5	0.7	0.3	3.5
Correctly classified instances (percentage)	89.9	76.9	92.8	92.9	92.7
Kappa	0.000	0.217	0.468	0.476	0.476

Table 4 High level performance analysis for a multiclass prediction.

Method	Cor	Del	Ins	Precision	Recall	F-measure	SER
Naive Bayes	1467	7070	2432	17.2	37.6	23.6	243.7
Logistic Regression	1233	287	2666	81.1	31.6	45.5	75.7
CART	1268	248	2631	83.6	32.5	46.8	73.8
MultiLayer Perceptron	1282	386	2617	76.9	32.9	46.1	77.0

Table 5 Detailed performance analysis for a multiclass prediction.

values presented in the last two rows.

Table 5 presents detailed performance values for each one of the approaches, revealing that CART should be the best choice for this type of problem. It is also interesting to notice that while the best precision is achieved using a CART, the best recall is achieved using the Multilayer Perceptron. For this experiment, Logistic Regression presents the second best performance, but all metrics reflect this difference coherently.

6.2.1 Detailed CART Results

Taking into account the results previously presented, the following results are achieved using CARTs. Table 6 presents the best results achieved for automatically identifying each one of the structural elements that are related with disfluencies. The table reveals that, from all the structural elements related with a disfluency, the interregnum is by far the easiest to detect. However, that is due to the fact 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. The second best results considering both the F-measure and se SER are 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 high and characteristic contrasts in terms of feature values. Detecting the repair zone can also be performed at a considerably

	Cor	Del	Ins	Precision	Recall	F-measure	SER
IP	379	145	600	72.3	38.7	50.4	76.1
interregnum	684	7	0	99.0	100.0	99.5	1.0
other word inside disfluency	42	35	1168	54.5	3.5	6.5	99.4
repair	163	61	863	72.8	15.9	26.1	90.1
outside disfluency	34425	2492	109	93.2	99.7	96.4	7.5
Overall performance	35693	2740	2740	92.9	92.9	92.9	14.3

Table 6 Zone discrimination CART results.

Classified as →	IP	interregnum	in-disf	repair	outside disf
IP	379	0	15	8	577
interregnum	0	684	0	0	0
other word inside disfluency	95	0	42	16	1057
repair	2	0	3	163	858
outside disfluency	48	7	17	37	*34425*

Table 7 Cart confusion matrix.

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 not the interregnum, which correspond to words that most of the times are in fact fluent. The previous analysis can also 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 being 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. Our experiments on the automatic identification of disfluent sequences suggest that similar results can be achieved using either CARTs and Logistic Regression. While CARTs tend to favor a better precision, Logistic Regression conducts to a better recall. Our experiments that distinguish between structural elements related to disfluencies suggest that CARTs are consistently better than the other tested approaches. In terms of computational effort, Logistic Regression is the best choice, being more than 10 times faster than Naive Bayes and around 100 times faster than Multilayer Perceptron.

This paper complements the first studies that have been performed on detecting disfluencies and disfluency related regions for portuguese University Lectures [13]. For the future, we are planning a similar work for distinguishing between disfluency locations and punctuation marks.

References

1 F. Batista, H. Moniz, I. Trancoso, N. Mamede, and A. I. Mata. Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. *Journal of Speech Sciences*, (3):–, 2012.

2 Cheng-Tao Chu. Detection of word fragments in mandarin telephone conversation. In *INTERSPEECH*, 2006.

3 H. H. Clark. *Using language*. Cambridge University Press, 1996.

4 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.

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

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

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

8 Yang Liu. Word fragment identification using acoustic-prosodic features in conversational speech. 2003.

9 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.

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

11 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 Transaction on Audio, Speech and Language Processing*, 14(5):1526–1540, 2006.

12 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.

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

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

16 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.

17 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.

18 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.

19 J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro. Broadcast news subtitling system in Portuguese. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1561–1564, 31 2008-April 4 2008.

20 D. O'Shaughnessy. Recognition of hesitations in spontaneous speech. In *IEEE Conference on Acoustic, Speech, and Signal Processing*, pages 521–524, 1992.

21 T. Pellegrini, H. Moniz, F. Batista, I. Trancoso, and R. Astudillo. Extension of the lectra corpus: classroom lecture transcriptions in european portuguese. In *GSCP 2012*, 2012.

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

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

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

25 Elizabeth Shriberg, Rebecca Bates, and Andreas Stolcke. A prosody only decision tree model for disfluency detection. In *Proc. EUROSPEECH*, 1997.

26 M. Snover, B. Dorr, and R. Schwartz. A lexically-driven algorithm for disfluency detection. In *Proceedings of HLT Conference/ NAACL annual meeting*, 2004.

27 Don Baron Elizabeth Shriberg Andreas Stolcke. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In -, 2002.

28 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.

29 A. Veiga, S. Candeias, C. Lopes, and F. Perdigão. Characterization of hesitations using acoustic models. In *International Congress of Phonetic Sciences - ICPhS XVII*, 2011.

30 Andreas Stolcke Yang Liu, Elizabeth Shriberg. Automatic disfluency identification in conversational speech using multiple knowledge sources. In *INTERSPEECH, ISCA, (2003)*, 2003.

31 Jui-Feng Yeh. Speech recognition with word fragment detection using prosody features for spontaneous speech. 2011.

32 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.