ISCTE IUL
University Institute of Lisbon

School of Technology and Architecture

Department of Information Science and Technology

Automatic Detection of Disfluencies in a *Corpus* of University Lectures

Henrique Rodrigues Barbosa de Medeiros

Dissertation presented in partial fulfillment of the Requirements for the Degree of *Master in Computer Science Engineering*

Supervisor:

Prof. Doctor Fernando Manuel Marques Batista, ISCTE-IUL & INESC-ID

Co-Supervisor:

Doctor Helena Gorete Silva Moniz, INESC-ID & CLUL

September, 2014

Abstract

This dissertation focuses on the identification of disfluent sequences and their distinct structural regions. Reported experiments are based on audio segmentation and prosodic features, calculated from a corpus of university lectures in European Portuguese, containing about 32 hours 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. The best results concern the detection of the interregnum, followed by the detection of the interruption point. Several machine learning methods have been applied, but experiments show that Classification and Regression Trees usually outperform the other methods.

The set of most informative features 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.

We have also conducted initial experiments on automatic detecting filled pauses, the most frequent disfluency type. For now, only force aligned transcripts were used, since the ASR system is not well adapted to this domain.

This study is a step towards automatic detection of filled pauses for European Portuguese using prosodic features. Future work will extend this study for fully automatic transcripts, and will also tackle other domains, also exploring extended sets of linguistic features.

Resumo

Esta tese aborda a identificação de sequências disfluentes e respetivas regiões estruturais. As experiências aqui descritas baseiam-se em segmentação e informação relativa a prosódia, calculadas a partir de um corpus de aulas universitárias em Português Europeu, contendo cerca de 32 horas de fala e de cerca de 7,7% de disfluências.

O conjunto de características utilizadas provou ser discriminatório na identificação das regiões contidas na produção de disfluências. Os melhores resultados dizem respeito à deteção do interregnum, seguida da deteção do ponto de interrupção. Foram testados vários métodos de aprendizagem automática, sendo as Árvores de Decisão e Regressão as que geralmente obtiveram os melhores resultados.

O conjunto de características mais informativas para a identificação e distinção de regiões disfluentes abrange rácios de duração de palavras, nível de confiança da palavra atual, rácios envolvendo silêncios e declives de *pitch* e de energia. Características tais como o número de fones e sílabas por palavra provaram ser mais úteis para a identificação do interregnum, enquanto pitch e energia foram os mais adequados para identificar o ponto de interrupção.

Foram também realizadas experiências focando a deteção de pausas preenchidas. Por enquanto, para estas experiências foi utilizado apenas material proveniente de alinhamento forçado, já que o sistema de reconhecimento automático não está bem adaptado a este domínio.

Este estudo representa um novo passo no sentido da deteção automática de pausas preenchidas para Português Europeu, utilizando recursos prosódicos. Em trabalho futuro pretende-se estender esse estudo para transcrições automáticas e também abordar outros domínios, explorando conjuntos mais extensos de características linguísticas.

Keywords

Automatic disfluency detection

Spontaneous speech

Corpus of university lectures

Machine learning

Speech processing

Prosodic features

Filled pauses

Statistical methods

Palavras Chave

Deteção automática de disfluências

Fala espontânea

Corpus de aulas universitárias

Aprendizagem automática

Processamento de fala

Features prosódicas

Pausas preenchidas

Métodos estatísticos

Agradecimentos

Esta tese não teria sido possível sem o suporte de vários intervenientes. É meu desejo agradecer ao INESC-ID e ao ISCTE-IUL por me terem acolhido de forma impecável. Quero também agradecer ao meu Orientador, Professor Doutor Fernando Manuel Marques Batista, e Co-Orientadora, Doutora Helena Gorete Silva Moniz, por me abençoarem com suporte e compreensão. Muito obrigado por me ajudarem a progredir tanto pessoal como profissionalmente. Gostaria também de agradecer ao Professor Doutor Luís Miguel Martins Nunes por me ter dado oportunidade e por ter motivado o meu gosto pela área de aprendizagem automática pela forma excelente como conduziu as aulas.

Quero, ainda, agradecer ao meu pai, Mário Henrique Barbosa de Medeiros, por todo o apoio que me prestou e por ter estado sempre do meu lado.

Lisboa, Setembro de 2014

Henrique Rodrigues Barbosa de Medeiros

“I don’t think... then you shouldn’t talk, said the Hatter.”

(Lewis Carroll, Alice in Wonderland)

Contents

List of Figures

List of Tables

1 Introduction

This dissertation focuses transversely on the automatic detection of disfluencies, a natural occurring speech phenomena that is used for the online editing of information. These structures represent a challenge in developing speech processing systems (ASR), mostly because they interact with the detection of surrounding words and sentence boundaries, interfering with overall error rates.

In this dissertation research aims at assessing the best suited cues and methods for automatically discriminating such structures, the distinct regions of disfluency, and filled pauses (FPs) in particular for European Portuguese. Additionally, the best results achieved in the FP classification experiments are compared to the in-house ASR solution, implemented by adding possible phonetic disfluent sequences to the lexicon of the recognizer. The corpus used is representative of the university lecture domain, being mostly composed of spontaneous speech. This specific domain has received increased attention lately, due to the potential for applications such as automatic multimedia content generation. Such application could support hearing-impaired students, and also improve the students learning experience, by providing automatically transcribed material of the course, potentially aiding both in present or distant learning. Additionally, this domain is similar to the one present on talks or public speeches, containing potential for the production of applications also in these areas.

The feature set used in the classification experiments is based on prosodic and lexical characteristics, but the overwhelming majority relies on prosody. Prosody can make ASR systems more robust when lexical information is not reliable, and also in the opposite case, since it provides valuable information for distinguishing disfluency types and inner-regions. The methodological approach is based on state of the art supervised machine learning algorithms, comprising generative and discriminative classifiers: Classification Trees (CART, J48), an Artificial Neural Network (Multilayer Perceptron), Logistic Regression, and a Naïve Bayes.

Research described here aims at answering the following questions: what features contain the most relevant information for detecting disfluent structures in general? what are the most relevant features for the detection of each disfluent zone, and what is the degree of difficulty associated to each of them? what are the most relevant features for detecting FPs? and, how good is the achieved performance in comparison to the achieved in-house solution baseline. Additionally, we also wanted to observe the

impact of including filled pauses and fragments as features (the most common disfluency genres) on the classification experiments here performed. Doing so allows both to explore the impact of such information on the classification of disfluencies, and also to determine the impact of the proposed set of acoustic features on the performed classification tasks. Additionally, most results in an initial phase are exposed both for automatic ASR transcriptions, and also for corresponding orthographically corrected versions of the material, allowing to verify the impact of using ASR transcriptions on the suggested classification experiments. The impact of filled pauses and fragments is verified on both setup cases. The results achieved represent a step forward in the understanding and automatic detection of disfluent phenomena for European Portuguese. Although reckoning the importance of understanding both how FPs and FRGs behave, as well as that of the remaining disfluent genres is important, in this dissertation only the FP genre is targeted in particular.

1.1 Disfluencies

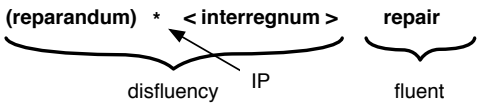

Figure 1.1: Disfluency regions.

Disfluencies are a natural linguistic phenomena that interrupts the normal flow of the oration, being used in one way or another by almost every speaker. These structures originate from several competing cognitive processes: thinking what to say; how to say it, and; coordinate actions with others, frequently resulting in non-linearity in speech (Swerts et al., 1996; Clark and Tree, 2002; Levelt, 1989). The effects manifest through breaks, irregularities, or non-lexical vocables that arise from fluent speech, which can either pass without notice or hamper message understanding, depending on the intensity at which they are uttered. There are several degrees of disfluency, ranging from not noticeable, to making the speaker hard to comprehend.

Disfluencies have a structure composed of several possible regions, *vide* Figure 1.1. The first region represents the zone to be auto-corrected, the reparandum. At some point the speaker detects a problem, and according to a "Main Interruption Rule" halts the production process, resulting in an interruption point (IP) (Levelt, 1983). Follows an optional editing phase or interregnum, filled with expressions such as "uh" or "you know", silent pauses are also frequent (Levelt, 1983; Nakatani and Hirschberg, 1994; Shriberg, 1994; Levelt, 1989). Finally a repair region, where speech fluency is recovered. All these regions are sequential in nature, and the true disfluent material relies solely on the reparandum and editing phase, although some cues to disfluency may also be tracked in the repair region. Regarding the typology of such events, according to Shriberg (1994) these can be: filled pauses (FPs), prolongations, repetitions,

deletions, substitutions, fragments (FRGs), editing expressions, insertions or complex sequences (more than one category uttered).

Disfluencies are studied from various angles (Psycholinguistics, Linguistics, Text-to-speech, Automatic Speech Recognition), languages and settings (human-human and human-computer), depending on the level of planning (spontaneous *vs*. prepared) and perspective (theoretical or applied) (Shriberg, 1994; Rodriguez and Torres, 2006; Dufour et al., 2009; Moniz et al., 2009). The studies comprised in Fromkin (1971); Heike (1981); Levelt (1989); Allwood et al. (1990) argue that disfluencies are not simply erratic waste left to chance when the speaker interrupts the structuring of his message, pointing the existence of linguistic patterns with systematic regularities. Disfluencies may sometimes aid in the process of conversation, accounting for tasks such as maintaining / obtaining the conversation turn, or introducing a new topic in conversation (Gravano et al., 2011). The most common disfluency types relate to filled pauses, which are used as mechanisms to take or hold the floor (Hieke, 1981; O'Connell, 2005). Some studies support the elimination of disfluencies in order to obtain the desired message by the speaker in the cleanest possible form (Honal and Schultz, 2005; Liu et al., 2006), while other studies demonstrate that disfluencies should be transcribed and integrated, so that the message can be the accurate portrayal of what the speaker has produced, and can therefore illustrate the specificity of the communicative situation, the strategies of the speakers, the dual role being both a transmitters and receivers of messages, or even the emotional states of the interlocutors and the dynamics of communicative interactions Benus et al. (2006); Adell et al. (2008); Parlikar et al. (2010). Although distinct, these perspectives share the consciousness that disfluencies display regular patterns that respect distributional rules, can adapt to the adjacent linguistic material at different levels of linguistic analysis, and behave as structuring regulator mechanisms for online planning.

It is known that the particular types of disfluency are unrelated to the corresponding rates in speech, or localization in the speech segment, meaning that speakers display individual strategies. It is also known in the literature that disfluency regions have idiosyncratic acoustic properties, that allow to distinguish between genres (Hindle, 1983; E. Shriberg, 1999; Shriberg, 2001), and the same applies for distinguishing disfluent regions. The most relevant properties of disfluencies happen in the reparandum and editing phases, but some effects can also be found in the repair. This dissertation focusses on identifying these regions based on their distinct properties, the whole disfluent region, and FPs.

Two general trends seem to exist for disfluency: containing material only in the editing phase, such as FPs; or on the other hand containing material in the reparandum and repair. Shriberg (2001) presents evidence suggesting that disfluencies are related to factors associated with the speaking context. The following section describes how these structures relate to the research objectives.

1.2 Motivation and Goals

Despite significant recent advances many speech recognition systems still produce an output that is hard to read and to treat by automated tools for natural language processing. The result generally resembles a sequence of words where no detection of structural phenomena is considered (Consortium, 2004). Disfluencies in particular mitigate the reading and processing of results, disrupting normal speech execution, while producing elements that may hamper understanding (Jones et al., 2003; Heeman and Allen, 1999). Maximizing disfluency detection is vital in the production of robust ASR systems (Stouten and Martens, 2004; Dannélls, 2007). These structures disrupt the normal course of the sentence and, when for instance word interruptions are concerned, they also give rise to word-like speech elements which have no representation in the lexicon of the recognizer. Since disfluencies characterize speech, although displaying more accented contours in spontaneous speech, considering them becomes essential in speech recognition systems (Liu et al., 2006).

An automatic speech recognition system (ASR) is often a pipeline with several modules, where each one feeds the subsequent with more levels of information. Disfluencies are known to have impact in the ASR modules, since they are frequently miss-recognized and may also lead to the erroneous classification of adjacent words, increasing the word error rate (WER). Additionally, predicting disfluencies permits a more clear filtering of the information, providing the base for structural and semantical analysis. A trustworthy prediction of such events, would allow to disambiguate between sentence-like units, and also between those units and disfluency boundaries (Liu et al., 2006). Detecting disfluencies, as well as performing the segmentation and capitalization tasks, are relevant MDA (Metadata Annotation) tasks.

The study described in Kahn et al. (2004) stresses the importance of the detection of disfluencies and sentence boundaries in the process of automatic segmentation, highlighting the importance of accounting for structural metadata in the subsequent development of these systems. These sequences have a huge impact on several post-processing tasks: language modeling, since disfluencies may occur in clause or phrase boundaries; speech characterization, whether it is planned or spontaneous; speaker characterization; production of multimedia content; speech summarization; automatic captioning; automatic translation; production of multimedia content. Systems that use ASR and text-to-speech (TTS) also benefit from predicting disfluencies, and also in the case of TTS alone, accounting for more natural imitations.

Nakatani and Hirschberg (1994) shows that detecting disfluencies is not a trivial task, additionally other studies found combinations of cues that can be used to identify disfluencies and repairs with reasonable success (Clark, 1996; Hindle, 1983; Goto et al., 1999). It is known that some zones present better cues for detecting the disfluent region. Clark (1996); Hindle (1983); Levelt (1983) point characteristics for several disfluent types relating to segment duration, intonation characteristics, word completion,

voice quality alternations, and pattern coarticulations.

It is known that fragments (FRGs) can be problematic for recognition if not considered and fairly identified (Yeh and Yen, 2012; Yang Liu, 2003), but they are also referred to as important cues to disfluent regions identifiable throughout prosodic features. Liu (2003) shows that FRGs can present different significant characteristics across languages. Filled pauses (FPs) are also problematic since they can be confused and recognized as small functional words, resulting in structures that decrease the ASR performance. It is known that, even though the phenomena varies between languages, there are constant similarities and pattern trends. Several studies point clues for detecting disfluencies such as: sentence length, modifications in segment durations, intonation, voice quality, vowel quality and coarticulation patterns, the presence of other disfluencies in the sentence, combinations of both these features both across and within speakers, word related features of disfluency, the rate of the cut-off words and the rate of editing phrases (Shriberg, 1994, 2001).

Studies such as Gormana et al. (2000) show that speakers are capable of perceiving disfluency in unknown languages, and that specific acoustic and temporal cues aid in disfluency perception. The results presented show that the duration features robustly explain differences in the perception of disfluency, and that pitch contours are correlated with differences in disfluency perception for different languages. Lai et al. (2007) describes a cross-linguistic perception experiment in which FPs and partial words are tested in terms of human perception, between English, German and, Mandarin. The results show that the subjects can distinguish between fluent and non-fluent events at a level above chance, pointing the existence of cross-linguistic phonetic cues. Additionally the results show that FPs were easier to identify than partial words (fragments).

Currently, INESC-ID develops AUDIMUS (Meinedo et al., 2003; Meinedo, 2008), a speech processing system that could benefit from an improved disfluency detection approach. Such module would represent an alternative information treatment, that could improve both textual legibility and the automatic scoring system's performance (Jones et al., 2003; Kim and Woodland, 2001; Heeman and Allen, 1999). Although the ASR system has to account for all the disfluent categories: FPs, prolongations, repetitions, deletions, substitutions, FRGs, editing expressions, insertions, and complex sequences, in this study the focus relies on the detection of regions related to disfluency, the disfluent structure as a whole, and FPs in particular.

This dissertation represents a step forward towards the characterization and automatic detection of disfluencies for European Portuguese. The proposed approach is based on machine learning methods, which are known to produce robust results, instead of just adding possible disfluent phonetic sequences to the ASR lexicon. Research aims at investigating the most expressive set of features for each target class, and testing solutions for pattern recognition. The data set available is a disfluency rich corpus of college classes, containing approximately 32 hours of speech and 7000 different disfluency occurrences

(7.6%) (Trancoso et al., 2008). Corresponding automatically and manually annotated data is used as a information source for both training and evaluation. Since the ASR system is not well adapted to the specific domain, which is rich in spontaneous speech, both force aligned and raw transcripts are used.

The research questions are the following: how does an acoustically based approach and machine learning techniques perform in comparison to the in-house implemented solution, that is based on the introduction of disfluent elements in the ASR lexicon? what features contain the most relevant information for detecting disfluent structures in a general way? what are the most relevant features for the detection of each disfluent zone and what is the degree of difficulty associated to each of them? what are the most relevant features for detecting FPs and how is the achieved performance in comparison to the achieved in-house solution baseline? Additionally, initial experiments access the impact of filled pauses and fragments, which are the most common occurrences, for both data modalities (recognition / forced alignment).

The ASR AUDIMUS performs automatic identification of FPs with the aim of filtering and including rich transcripts for broadcast news. The filtering process was achieved by identifying speech regions with plateau pitch contours and energy values. The inclusion process was exclusively based on the integration of FPs in the lexicon with alternative pronunciations. Using known phonetic sequences of disfluencies to discriminate the phenomena makes the system intrinsically less robust to variations, and also harder to adapt to new speech domains (He and Young, 2004). The experiments reported here are a step forward in the prediction of FPs by means of encompassing a broader set of acoustic features, and also by testing distinct classification methods to evaluate the best performance achieved. The best results are obtained for the detection of FPs while using J48, corresponding to about 60% precision, and 61% f-measure. The proposed approach is compared with the one currently in use by the in-house speech recognition system, and promising results are achieved.

The following section describes the methodologies adopted from the literature.

1.3 Methodology

This subsection describes the methodological approaches adopted in this dissertation. The process of identifying potential features was undertaken based on the study of Portuguese as well as foreign languages, as there is evidence towards the existence of similar inter-linguistic phenomena related to disfluencies. It is known in the literature that lexical features have a greater influence on classifications than the ones that rely on prosody (O'Shaughnessy, 1994; Nakatani and Hirschberg, 1994; Heeman and et al., 1994; Bear et al., 1992; Shriberg et al., 1997). These studies present reduced robustness in cases where lexical information is not trustworthy, however, these also demonstrate that prosody is

helpful if constrained by the lexical information. Liu et al. (2006) shows that the combination of acoustical and prosodic information leads to improved results, while performing experiments on telephone speech, and testing 3 machine learning approaches: Maximum Entropy Models, Conditional Random Field, and Hidden Markov Model. This study reports a general superiority of discriminative over generative models. It is also known in the literature that removing disfluencies from the n-gram context contributes to reduce language model perplexity (Stolcke and Shriberg, 1996). Studies such as Moniz et al. (2012b) analyzed acoustic elements such as pitch, duration and, energy, to identify and distinguish between the various types of disfluent moments. Prosody can be tracked in the speech signal using phrasing units, words, syllables and phones. As mentioned in Moniz et al. (2012a) there are two main strategies widely used in the literature for detecting the distinct possible regions inherent to disfluent structures, while using prosodic features: (i) a contrastive strategy between the reparandum and the repair of fluency, manifested by pitch and energy increases at the onset of the repair; and (ii) a parallel prosodic strategy between this same areas, meaning, the repair mimics the tonal patterns of the reparandum (Levelt, 1983). Hindle (1983) suggests the existence of an edit signal capable of denouncing an upcoming repair. Manifestations of this signal can be tracked based on fragments (FRGs), repetition patterns, glottalizations, co-articulatory gestures and, voice quality attributes, such as jitter (perturbations in the pitch period) in the reparandum. Additionally it is also edited by means of significantly different pause durations from fluent boundaries, by specific lexical items in the interregnum, and via pitch and energy increases in the repair. For the task of detecting disfluencies the main focus thus becomes to detect the interruption point, or the frontier between fluent and disfluent speech, due to the inherent potential for discriminating the phenomena. However, in the excerpt used in this dissertation, the percentage of disfluencies that contain and interruption point is of 34.9% in the training set, and 35.2% in the test set, showing that these represent less that half the totality of disfluencies.

Since statistically based techniques are known to be more robust for modeling variations of the phenomena than a rule-based approach, this study tests such implementation. Experiments were conducted using the data-mining and machine learning software, Weka [1]. In the present work, the classification tasks are performed by means of supervised machine learning approaches, comprising generative and discriminative classifiers. Supervised methods use one or more inputs (x) and desired outputs (y), to learn a general rule for mapping from x to y. Generative models learn the joint probability distribution $p(x, y)$, while discriminative models learn the conditional probability distribution $p(y|x)$. Three discriminative models are tested: Logistic regression (LR), Classification and Regression Tree (CART), J48 and, Multilayer Perceptron (MP). As for generative models only Naïve Bayes (NB) was tested. All these methods are widely used in the literature, representing state of the art approaches. There are several machine learning approaches that have performed well on previous work, such as Conditional Random Field (CRF), which where not tested in the experiments described in this dissertation. Although not

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

experimented in the present work, these constitute potential targets for future experiments.

The ASR AUDIMUS (Meinedo et al., 2003; Meinedo, 2008) was trained for the Broadcast News domain, for that reason, it presents a word error rate (WER) of about 50% while facing university lecture recordings. 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, simulating perfect recognition conditions, in order not to bias the study with the poor results obtained with an out-of-domain recognizer. The set of features proposed is mostly composed of prosody, which is automatically extracted from the audio signal, and encoded into self contained XML files. Some of these files are then enriched by human annotators with information such as, sentence boundaries, disfluencies, disfluent genres, disfluent regions, etc, providing targets that allow to study and use such information. For sake of comparison, several results in this dissertation will be reported for both force aligned and automatic transcripts. To analyze feature influence the following processes were used: the top-most branches of the Decision Trees; Logistic Regression model weights, since these offer insight on feature influence as further reported in the next section.

2 Related Work

This chapter performs an overview of relevant literature for this work, comprising either European Portuguese related material, and also foreign languages related studies, given that compelling evidence exists towards the existence of cross language similarities in disfluencies. Research aims at exploring well suited methods and features, providing an overview of the state of the art on detecting disfluencies. Some studies overview other tasks in conjunction with detecting disfluencies, such as automatic segmentation, since the same feature genres are also used to describe such data, although in the present chapter only the perspective of disfluency is overviewed.

It is shown that the perception on disfluencies has suffered several mutations over time, going from a notion of mere speech errors to important communication devices, that aid in the conversation process, being used for several communicative functions. This notion however is still not widely accepted. Additionally, disfluency is reviewed in therms of the subjective human notion of behaving as fluent speech, showing these are recurrently considered by humans as fluent communicative devices. Some studies described focus on linguistic descriptions of the disfluencies, while others focus on the disfluency classification task, denoting more or less emphasis on the methodic, or feature representation strands of the classification routine.

The features types suggested for disfluency detection relate to, energy, word confidence, backchannel information, morphosyntactic information, etc. Several speech genres are included in the described studies, however in this chapter the emphasis relies on spontaneous speech and school classes. Domains such as telephone conversation, or task oriented dialogue are also target of overview, given that these are rich in disfluencies. Some studies focus the interruption point and interregnum phase, while other focus more on the repair. There are also studies that explore the relation between the disfluent material and the surrounding fluent content. Additionally some studies comprise results performed on mixed domain speech data.

The next sections are structured as follows: (i) overviews disfluencies under European Portuguese, (ii) the focus is on filled pauses (FPs) under European Portuguese, (iii) overviews disfluencies for other languages, (iv) describes FP detection directives found in other languages, (v) performs a summary of the literature overview performed in the whole chapter.

2.1 Overview of Disfluencies in European Portuguese

This section describes related work on the topic of disfluency for European Portuguese. The study by Moniz et al. (2011b) seeks to deepen the analysis of prosodic properties of disfluencies using the same source used in this dissertation (LECTRA Trancoso et al. (2008)), and a corpus spoken by Adolescents in School Context (CPE-FACES) Mata (1995). The time samples taken for the performed experiments correspond to 2 hours for CPE-FACES and 1.5 hours for LECTRA. The sample collected from LECTRA presents a percentage of disfluency of roughly 7%, in line with the value predicted by Shriberg (2001) of 5 to 10%. Two classification experiments are conducted using CARTs, and additionally an analysis over 10 hours of LECTRA is performed, characterizing disfluency both generally and in terms of distinguishable traces between disfluent categories. Both classification experiments are based on an identical setup, and the data is divided into training, validation and testing (60%, 20% and 20%, respectively). The classification experiments target the characterization of disfluencies in general, and also the vicinity of disfluency, in an attempt to capture disfluency related traces. For discriminating in a binary fashion fluent from disfluent, all levels of annotation are taken into account: the judgments of fluency / disfluency (as the main question), the prosodic information, morphosyntactic information, the speakers and the communicative situation (prepared speech and spontaneous).

In order to analyze the disfluent properties, the authors resort to the ToBI system[1], in order to asses the degrees of cohesion between words. In this system the breaks are annotated in accordance to a scale varying between 0 (maximum degree of cohesion) and 4. The authors report that the relevant occurrences range from 2 to 4. The 2p index is used for temporal rupture occurrences (full or virtual breaks associated with an intermediate prosodic constituent boundary), but the absence of tonal break. In turn, the rupture index 4 matches intonational boundaries of major constituents, characterized by strong temporal and melodic breaks, being associated with endings (full-stops, commas, question marks, exclamation marks).

In a fist binary classification experiment, the respective results of the judgments of fluency *vs*. disfluency indicated that prosodic cues are the most salient information, even more than the morphosyntactic information, enabling the differentiation of events based on two essential partitions in the CART: the phrasing and the prosodic contours. The boundaries of disfluency *vs*. fluency in the speech array prove to be important locations to look for traces. As for the CART partition relative to phrasing, the authors report that disfluencies produced after a prosodic constituent of 3 or 4 are majorly classified as fluent, while the ones that are produced inside a constituent with break indices of 2p are majorly classified as disfluent (78.3%). It is argued that the percentage of disfluencies occurring inside a constituent and still classified as fluent (21.7%) can be explained by two distinct behaviors: they are events produced in

[1] The ToBI system is a set of conventions for the task of transcribing and annotating speech prosody. The resulting transcription minimally indicates the intended prosodic grouping of an utterance and the corresponding tonal events.

the initial position of a constituent, with pitch reinicialization (10.4%); they are events produced in the final of a constituent with frontier tones that express discursive continuation, or even in the end of a constituent with termination contours. The authors argue that this behavior is expectable in the european portuguese neutral declarative phrases. The second most important CART partition (intonational contours) evidences that the events are produced in either 3º or 4º constituent frontiers, with ascending plateau intonational contours, are considered fluent (90%), while the events produced in the same locations but displaying descending contours, or with globalization effects are considered disfluent (72%).

In a second experiment the setup is only changed by the removal of the most important feature, phrasing, which is motivated by the goal of studying features that are more easily detectable in an automatic fashion. The results of this second experiment show that, if the events are produced with plateau contours, or of rising intonation, and the morphosyntactic information indicates that the constituent corresponds to a phrase or sentence, they are significantly considered fluent (88.7 %). Also, if the events are produced with pitch reset, then they are considered fluent in 70.7% of cases, were-as if there is a reboot of pitch and pitch contours are descendant, or glottalization effects occur, events are significantly classified as disfluent (95.3% and 80%, respectively).

In both experiments the performance for detecting disfluencies was above the baseline (better than random guess), in the first experiment the error is of 29.05%, while in the second experiment 32.9% is obtained respectively (when considering the 6 most important leaves of the CART tree). Surprisingly, event duration was not an expressive feature, as predicted by Shriberg et al. (1997), this feature seemed to have impact only when considering the first 12 leaves, but still the values remain bellow 50%.

Finally, an analysis is performed over the 10 hours recovered from LECTRA, showing the importance of prosodic patterns for detecting disfluencies, and verifying that the prosodic properties allow the distinction between prosodic categories. As expected, the authors describe regularities on the prosodic properties of disfluencies (energy, duration and, pitch) for the different categories. Filled pauses (FP) denote the lowest medium pitch, minimum maximum pitch, minimum minimum pitch, maximum maximum energy and, maximum minimum energy, being among the most distinguishable categories. The contrast of all categories clearly discriminates two patterns: i) FPs, complex sequences and prolongations are the categories that most significantly differ in all parameters; ii) only the repetitions and substitutions are indistinguishable.

Different conclusions are drawn from the obtained results: i) both the FPs and elongations are sustained vocalizations, which may explain the distinctive characteristics that these present on all other events; ii) the complex sequences may resemble (at least in the onset) to other prosodic constituents, showing the highest levels of pitch); iii) repetitions and replacements can not be significantly distinguished, since both may be associated with emphasis and informational reinforcement. Another point that this work stresses, is that disfluencies are under the control of the speaker, and that these require

mastery to be effectively used, which reflects on the degree of using disfluencies in the speech.

Moniz et al. (2012a) studies the importance of various factors in characterizing prosodic disfluencies, particularly in the repair moment (immediately after a disfluent structure), based on an analysis performed over a subset of the LECTRA, comprising a total of 16 hours, 7 speakers, 110427 words, from which 3.46% of words are disfluencies. The authors attend to answer several questions related to the contrast between the disfluent region and the oncoming repair, namely: if the speakers exhibit contrast / parallelism strategies; which prosodic parameters characterize these strategies, and; if distinct degrees of speaking proficiency in the university lectures domain are related to the use of combined prosodic strategies. As in this dissertation, several features are calculated for disfluent words, and the 2 adjacent words: pitch*;* raw energy; normalized mean; median, maxima, minima, and standard deviation; as well as part-of-speech tags; number of phones; durations of phones; words, and; inter-word pauses. It is reported a general tendency to repair fluency using prosodic contrast making strategies, regardless of the disfluency genre, which is in discordance with the vision described in Levelt (1983), that reports exclusivity for the *corrections* disfluent genre. In this view it is reported that upon the return to fluency, increases of both energy and pitch are produced. The authors find different contrastive prosodic degrees in the different disfluent genres, denoting FPs as the most distinct genre in therms of pitch increase and durational contours, and repetitions in respect to energy rising patterns. Substitutions show similar significant pitch / energy increase differences on the onset of the repair, or even on the slope within the repair. As for tempo patterns, it is reported that the region to repair is longer than the repair itself, and there is a strong trend manifested in lengthy silences between these regions. However the speakers monitor this lengthiness effect differently. The results point out to domain specificities, and the distinct regions inherent to disfluencies are also uttered with distinct prosodic properties. It is also reported that the selection of the disfluency types and distribution is speaker dependent, and that the speakers contrast the disfluent regions of disfluency using the minimum context possible.

According to the results reported by Batista et al. (2012), the most representative features for describing disfluent regions as a whole are: pitch and energy slopes, the differences between the corresponding slopes, and the tempo characteristics of the distinct regions and of the adjacent silent pauses.

Moniz et al. (2012b) addresses latter analysis performed over LECTRA corpus with the purpose of characterizing disfluencies. The authors report several interesting facts: pitch and energy slopes are significantly different between the disfluency and the onset of fluency and also for the unit before the break point, that those features are also relevant to disfluency type differentiation and, they seem to bee speaker-independent. It is also reported that the best set of linguistic features found for predicting the onset of fluency are pitch and energy resets as well as the presence of a silent pause immediately before a repair. The results reinforce a contrast strategy rather than one based on parallelism. For all disfluencies there are pitch and energy increases in the repair, but for FPs, deletions, and repetitions,

the degrees of contrast created between this zone and the reparandum are quite distinct.

2.2 Overview of Filled Pauses in European Portuguese

For European Portuguese, studies that rely on the topic of disfluency focus mostly on filled pauses (FPs). However the nature of these studies relates more to the characterization of the phenomena, than to the perspective of automatic classification.

Moniz et al. (2007) comprises a study based on a corpus of prepared (non-scripted) and spontaneous oral presentations in high school context under European Portuguese. This study provides an opportunity to contrast disfluent phenomena in spontaneous and prepared speech under European Portuguese. The focus relies on studying the contextual / prosodic distribution and temporal patterns of FPs and segmental prolongations, as well as on the way those are rated by listeners. The relative rates of disfluency types is also explored. The Corpus extracted from the CPE-FACES Mata (1995) is composed of 10 oral presentations Mata (2000), and the used excerpt corresponds to 2 hours, 10 minutes and 3 seconds, translating to 11,851 words, 9,708 and 2,143 for the prepared and spontaneous presentations, respectively. The speakers are 1 female teacher of Portuguese, representative of first language teaching, and 4 students (2 male and 2 female). Disfluency annotation is marked according to Shriberg (1994), and additional annotation was added containing information relative to the syntactic and prosodic context of the disfluencies, as well as those of all the silent pauses in the corpus. A total of 1569 disfluencies are observed, resulting in a proportion of 13.24% disfluencies, a higher rate than the one predicted by Trancoso et al. (2008) for Portuguese or Candea (2000) for French. Three distinct forms of FPs were found in the CPE-FACES corpus: (i) an elongated central vowel only; (ii) a nasal murmur only; and (iii) a central vowel followed by a nasal murmur. Prolongations present the highest frequency rates, and may occur as single isolated events far more often than FPs. The results suggest that FPs and segmental prolongations occur in complementary distribution, being used as a device to both sustain fluency, and gain time before syntactic complex units. The authors report similarities in terms of behavior and functions between these categories, reporting that these may be considered in complementary distribution, obeying to general syntactic and prosodic constraints.

As for other languages the results are consistent with task and speaker dependencies in therms of characteristics and relative use of disfluencies. Filled pauses and prolongations differ from the remaining disfluency types, as they often occur as single, isolated events, while repetitions, substitutions, truncations, insertions and editing expressions tend to combine with each other forming complex sequences. Three facts are pointed out in FPs: (i) ”aam” generally occurs at major intonational phrase boundaries, (ii) “aa” is the most likely form at minor intonational phrase boundaries, even though it may occur in practically all contexts, as it is the only form used by two of the speakers; (iii) “mm”, is always cliticized onto

prior elongated words.

This study concludes that FPs may be viewed as as manifestations of planning effort at different levels of the prosodic structure, also suggesting that FPs and prolongations occur in complementary distribution and are used as a device to both sustain fluency and gain time before syntactic complex units.

Following the previous study, Moniz et al. (2008) aims at extending previous analysis to an enlarged corpus, in order to verify the consistency of previously assumed facts, namely: the listeners ratings of several kinds of disfluencies, if segmental prolongations occur more frequently than FP and are better rated by listeners. The authors report that, contrarily to other languages, FPs and prolongations are used in European Portuguese to signal upcoming delays, and to gain time before syntactic complex units, as instances of the same device occurring in complementary distribution. The corpus is the same as the previously described study, now comprising about 12 hours with annotations for disfluencies, from which 4 hours are annotated for fluency ratings (2 high school, 2 University), and the sentences are annotated for ease of expression as felicitous or infelicitous. On a scale from 1 to 5, when only average answers of 4 or 5 were considered felicitous, 3 different trends of disfluency phenomena emerge, which are associated with different acceptability rates: (1) FPs and prolongations on top; (2) then substitutions and deletions; (3) fragments (FRGs), repetitions and complex disfluent sequences are considered less felicitous. Prolongations and FPs occurring in felicitous moments are regularly scaled relatively to their adjacent constituents, but this is not the case for repetitions and FPs occurring in infelicitous moments. In this view, for single FPs and prolongations, the presence of a silent pause preceding the repetition appears to be crucial, since it's removal strongly induces negative judgments.

The work described in Moniz et al. (2009) represents further work on exploring the prosodic cues of disfluencies under European Portuguese. A classification experiment using CART is performed, in an attempt to discriminate the most salient prosodic features for the task of classifying disfluencies as either fluent or disfluent, as considered by listeners. The corpus used is a subset of CPE-FACES and LECTRA, comprising 15h for the former (two teachers and twenty five students), and 10h for the latter (five teachers) (Trancoso et al., 2008). Manual annotations for disfluencies and disfluency ratings were added to subsets of these corpora, 2h for the CPE-FACES, and 1.5h for LECTRA. The corresponding disfluency rate is 13.24% (1569 disfluencies and 11,851 words) in the CPE-FACES sample, and 3.16% (273 disfluencies and 8636 words) on LECTRA. For the classification experiment the data is divided into training, validation and test data (60%, 20% and 20%, respectively). The set of features used for the experiment are: (dis)fluent judgements (as target feature), disfluency type, break indices, pitch *(f_0)* contour, f_0 restart, morphosyntactic information of the adjacent words, morphosyntactic information of the disfluency, speaker and speech situation (spontaneous and prepared non-scripted speech). The test misclassification rate was 29.05%, and 56.4% of disfluencies where considered disfluent and 43.6% as

fluent.

Since the results of the previous experiments where consistent with those of Moniz et al. (2007), in the sense that pointed out the importance of break indices and phrasing in fluency judgements, an additional study is performed to evaluate the prosodic constituents of the stimuli. The study focuses on segmental prolongations, FPs and repetitions, all categories previously associated to planing efforts (Clark and Tree, 2002; Clark and Wasow, 1998). The results point that the FPs judged fluent are uttered in a tonal space in-between the prosodic adjacent constituents, have stationary f_0 contours, and behave mostly as parentheticals. When FPs are considered infelicitous, they are produced in a lower register with descending contours, disrupting tonal scaling. It is reported FPs tend to occur between the previous brake and the ongoing of the conversation, and are uttered at a tonal space in between adjacent prosodic constituents.

In sum it has been demonstrated that prosodic phrasing is of crucial importance for the task of classifying the perception of disfluency, and also that contour shape is also important. The results support the view that disfluencies may behave and even be rated as fluent devices, and that speakers control different segmental and suprasegmental aspects when producing disfluencies, which in many occurrences seems to happen in a surgical way, adequately adjusting to the adjacent constituents.

The work described in Veiga et al. (2011) explores the acoustic-phonetic properties of hesitation phenomena, in order to identify and annotate some of these events in a spontaneous speech corpus of Portuguese broadcast television news collected by the authors. The authors use a corpus collected from podcaster television news, comprising around 22 hours of non-annotated speech not representative of spontaneous speech, and not much populated by hesitations. Filled pauses and Hesitations are annotated in a semi automatic way using a phone recognizer with several restrictions in therms of phone sequences and durations, and posterior human revision is performed. The goal is to study the possible acoustic-phonetic cues to detect FPs and hesitations (the same as prolongations), such as pitch, energy, spectral and durational characteristics, as well as their relation with phones, in an attempt to contribute to improve the acoustic modeling for spontaneous speech recognition systems under European Portuguese, and to confirm the constancy of the acoustic parameters in comparison to other languages. For every event, average values for pitch and energy, as well its deviation are computed, along with equivalent values for spectrum, using 32 frequency bands (on a mel scale), and corresponding deviation from the average spectrum in the segment. Additionally, the event detector implements a confidence measure based on the phone durations in each occurrence. Sometimes the difference from FPs or hesitations is not obvious, and is distinguished only via phonetic context. The authors report that, extensions occur mainly in prepositions and on the last syllable. It is shown that, most of the times, hesitation segments present negative gradient values, decaying smoothly during hesitations. The variation produced is very small, the standard deviation of pitch is on average around 15 Hz and standard deviation of energy is on

average around 2.7 dB. The parameter based on standard deviation of spectral band energies shows a similar behavior. Additionally the authors argue that the pointed characteristics do not separate well FPs and hesitations, supporting the fact that distinguishing between these types is highly ambiguous without a context.

2.3 Overview of Disfluencies in Other Languages

The work described in Shriberg et al. (1997) aims at assessing the detection results of four disfluent categories: filled pauses (FPs), repetitions, repairs and, false starts, based on an acoustic model. The corpus used consists of ”mixed sex” telephone conversations, and the training set is composed of 500,000 words, reserving 60,000 for testing (12%). Several classification experiments take place, featuring a database of speech-based automatic and manual transcriptions of disfluent phrase boundaries, timestamps and raw acoustic measurements. The classification approach is based on CART-like decision tree classifiers. Results suggest that prosody presents a valuable knowledge source for the task of automatically detecting disfluencies in spontaneous speech. The classification is based almost exclusively on features extracted before, or at the zone immediately preceding a silence found between the reparandum and repair. The main features reported are length, the distance from a pause, and pitch (fundamental frequency). It is reported that, the relative use of these features was generally similar for the four different disfluency types. Using a prosody-exclusive model, better results than the baseline on the task of identifying FPs, repetitions, repairs and false starts, are achieved, denoting the superiority of the prosodic model for the task of detecting false starts given a correct transcription. Finally, results show that the combination of a prosodic model with a specialized language model overcomes the use of only one of these models.

The work described in E. Shriberg (1999) aims at understanding phonetic consequences of disfluency on changes in segment duration, intonation, level of spoken word completion, voice quality and, quality / patterns of vowel coarticulation. The authors highlight markings located on the reparandum and editing phase, the characteristics pointed for disfluency detection are: changes in the length of segments, features on intonation, word completion, voice quality, and quality of vowel coarticulation patterns.

Savova and Bachenko (2003) studies 4 disfluency types (repetitions, substitutions, replacements with repetitions and, repetitions with insertions) from an acoustic perspective, relating exclusively to intonation and duration. Experiments are conducted based on an English corpus, representative of semi-spontaneous speech. Results suggest that the detection of different disfluent types is only feasible thru the combined use of various prosodic characteristics. Research also point that the prosodic features vary depending on the type of discourse, identifying an interdependence between the characteristics of early repair (error correction), and the start / end of Reparandum (disfluent moment). Additionally this

study evidences that their offsets and prosodic features vary depending on the type of discourse.

The work described in Stolcke and Shriberg (1996) verifies the impact of excluding, in turn, FPs, deletions and, repetitions, prior to creating a language model. The authors build language models based on the n-gram philosophy, one considering disfluencies, and the others without considering disfluent information. Disfluencies are modeled in a tri-gram fashion. Based on large portions of the switchboard corpus, the language models are used to study the before-mentioned disfluent categories. The authors omit FRGs from the annotations, but argue that successful FRG recognition may serve as extra evidence for repetitions and deletions, as well as for other disfluent events. The perplexity of the surrounding words to the disfluency is also studied, in order to assure the model doesn't penalize these tokens.

Results show that the removal of FPs augments model perplexity, particularly at the following word. Filled pauses are, on average, the best predictor of the following word, and not the context preceding the FP. The achieved disfluency model reduces word perplexity on neighboring words, however, the number of disfluent events in the corpus is very reduced, and the impact is not notorious in the accuracy.

Another experiment takes place based on excerpts of the Switchboard corpus, annotated for clauses, and deleting only medial FPs. This experiment showed that FPs tend to occur at clause boundaries (between clauses), since perplexity on the word following FPs increased greatly. This supports the view that disfluencies should be considered jointly with the sentence segmentation task. Removing repetitions and deletions results in slightly lower perplexity than in the FP case, but the error rate suffers no influence. Contrarily to FPs, the complete removal of repetitions yields a positive impact in the surrounding context perplexity. The authors also test if the words following repetition might be better predicted by the repetition itself as in the FP case, or by the words themselves. In this case the words turn out to be better predictors.

The following words to deletions are also tested for predictability based on the disfluent class or the constituting words, but as for repetitions the words themselves are better predictors. This study showed the reduced potential of cleaning up disfluencies from the transcription, in therms of language model perplexity. The authors thus argue that modeling disfluencies solely on the language modeling probably will not improve the WER significantly, and suggest the use of acoustic and prosodic information to improve classification performance.

The study comprised in Shriberg and Stolcke (2002) presents a framework for automatically detecting structural phenomena in ASR outputs, combining prosodic and lexical information to perform several classification or tagging tasks, namely: sentence segmentation, disfluencies (interruption point detection), topic segmentation, dialog act tagging, overlap modeling, among others. The classification approach is mainly based on CARTs. Additionally, strategies for combining prosodic and lexical information are explored.

Prosodic feature extraction is done directly and in a fully automatic way, using a forced alignment of the transcripts to extract features for target classes, providing the model opportunity to choose the level of granularity of the representation that is best suited for the task.

A set of features based either on raw elements pitch *(f_0)*, pauses, segment durations, and energy), or derived features based on subsequent computations is acquired for the target classes. These are normalized in various ways, conditioned on certain extraction regions, or conditioned on values of other features, without the use of intermediate abstract phonological categories such as pitch accent or boundary tone labels. From the phone level alignments several temporal elements are obtained, namely: durations of pauses and various measures of lengthening (syllable, rhyme, and vowel durations) and speaking rate. Post processing allows to obtain the f_0 baseline, f_0 estimates, computation of pitch movements and contours over the length of utterances or individual words, or over the length of windows positioned relative to a location of interest (e.g., around a word boundary), energy-based features follow the same trend.

Regarding the decision trees, the authors report 2 problems: greediness, and sensitivity in the case of highly skewed class sizes. The authors deal with these problems in the following manner: (i) to overcome greediness a feature subset algorithm wraps the standard tree growing algorithm, which performs the task of eliminating detrimental features from consideration; (ii) To deal with highly skewed class sizes the authors resample the training data in order to achieve similar class distributions, allowing the subsequent comparison of classification models either quantitatively and qualitatively, and also allowing an adequate integration of these with the language models.

Among other experiments the authors apply the previously described framework to the detection of interruption points in an excerpt of the Switchboard corpus representative of spontaneous speech, targeting several disfluency types, namely: hesitations, repetitions, deletions. This task is performed in conjunction with the sentence segmentation task. The testing is performed using a the language model. The goal of the language model it to model the joint probability of classes, word classes, and words $P(W, S)$, using the training data, using $P(S|W)$ for testing on the test data.

Using reference material tagged with information such as interruption points and disfluency categories, the authors use the HMM-based model integration philosophy. In HMM-based integration, the probability of the features given the class and the word sequence is computed from the prosodic model, and then used as observation likelihoods in a HMM derived from the language model. The goal of the HMM is to encode the unobserved classification possibilities in its state space. Finally an association of these states and the prosodic likelihood allow the computation of a joint model comprising the information of word sequence, features and, classes. Then, the HMM can compute the posteriors of the probability of the class, given the prosodic features, and word sequence. A language model is calculated standardly from the training text, and is then used as a HMM in which the states correspond to the

unobserved hidden word-boundary events. The HMM then calculates de joint probability of the prosodic likelihood scores based on the boundary event, given the features and word sequence, and uses them in combination with the HMM states, in order to constrain the HMM tagging result on the prosodic features.

Results show that the combination of word and prosodic knowledge offered the best results, and point pause duration as the strongest feature. Additionally, the classifiers trained on the Switchboard corpus (spontaneous speech) material relied primarily on phone duration features. The authors also conclude that speech recognition accuracy can also benefit from prosody, by constraining word hypotheses through a combined prosody / language model.

The study described in Jones et al. (2003), is based on the premise that most disfluencies can be detected using primarily lexical cues, including characteristics that exclusively relate either to the word itself, or to the corresponding grammatical relation to the rest of the sentence. This study confirms the feasibility of obtaining disfluency detection results comparable to those of prosody based systems, using a lexically driven approach, without relying extensively on prosodic features. In order to verify this comparisons, several experiments are performed using two systems based on prosodic features. The data available is from phone conversations and television news, representative of both spontaneous and prepared speech, containing: marks for speaker (detection), sentence boundaries, editing disfluencies , fills and, breakpoints (zone between reparandum and repair). The aim is to classify each word as being a filler, an edition or a fluent element.

The learning strategy is based on ”transformation based learning” (TBL), a technique that relies on learning a referenced set of rules that transform an initial hypothesis with the purpose of reducing the corresponding error rate. The set of possible rules is found by expanding the imputed rule templates. The algorithm greedily selects the rule that reduces the maximum error rate, applies it to the data, and forwards the research to the next rule, stoping when no more rules can reduce the error rate below a certain threshold. The system output is an ordered set of rules, that can then be applied to test data in order to (in this case) annotate it with disfluencies.

Results are divided by reference manual, transcription, and speech domain. The highest error source concerns the presence of lengthy editing disfluencies. These elements resemble the repair region, having no apparent cues in the prosodic level, influential in detecting disfluencies. To this type, the suggested solution is based on the analysis of long distance dependencies, based on parsing, and semantically analyzing the text.

One of the most relevant and complete work that is still state-of-the-art in the field is described in Liu et al. (2006), this paper describes a system capable of detecting structural sentence limits, disfluencies and, *fillers* (filled pauses). The system combines information from several sources, including: lexical, and information from a prosodic classifier. This combination allows the detection of a set of disfluencies that would not be detectable using only one information type. The authors address the following set of

questions: i) what sources of information are useful in detecting various events, ii) what are the most effective approaches regarding statistical models in combining different sources of information iii) how is performance affected by several factors such as the domain of the data set, transcripts and event types, iv) if the extraction of metadata can be improved considering alternative hypotheses of words. Experiments use discriminative statistical models, including Conditional Random Fields (CRF), Maximum Entropy Models (ME) and a generative model based on Hidden Markov Model (HMM). Combinations of these models are applied in transcripts from television news (broadcast news) and as telephone calls, mainly characterized by spontaneous speech. The rating system is based on manually transcribed data, simulating perfect recognition results. Such methodology allows subsequent comparisons since the results of the classifiers depend on the performance of the recognizer transcription.

Results vary depending on the task, method and material used for training, revealing that discriminative statistical models generally outperform generative models. The results show that the use of the 3 listed models, together with lexical and prosodic features, represent the best approach towards obtaining maximum performance for disfluency detection, and that the use of several sources of textual information produces better results than the use of single-language models based only on words.

2.4 Overview of Filled Pauses in Other Languages

Literature on filled pauses (FPs) points out to several features used for predicting such events. O'Shaughnessy (1992) shows that FPs exhibit low pitch and plateau or falling tones. Shriberg (1994); Shriberg et al. (1997) evidence that FPs can be fairly detected using prosodic features related to duration, silent pauses, and pitch. The work of Goto et al. (1999) describes experiments on 100 utterances extracted from a Japanese spoken language corpus O. Hasegawa S. Hayamizu K. Tanaka K. Itou (1999). Based on small and constant pitch transitions and small spectral envelope deformations, this study achieves 91.5% precision and 84.9% recall on FP detection. Swerts et al. (1996) explore the interplay between FPs and discourse structure based on Dutch spontaneous monologues from a corpus of 45 minutes of speech containing 310 FPs. This work reports that stronger breaks in the discourse are more likely to co-occur with FPs than do weaker ones, that FPs at stronger breaks also tend to be segmentally and prosodically different from the other ones and they have more often silent pauses preceding and following them. Tsiaras et al. (2000) perform experiences on a corpus of Greek university lectures of approximately 7 hours containing 1124 occurrences of FPs. The authors report the utility of video information for improving precision and recall on the task of detecting FPs, achieving a precision rate of 99.6% and recall rate of 84.7%. This represents a considerable improvement over the 98.5% precision and 80.6 recall achieved using solely the audio stream.

2.5 Summary of the Literature Review

In sum, the literature overviewed in this section describes insights on the disfluent phenomena for several perspectives, languages, disfluency types, disfluency regions, as well as approach methodologies regarding the task of automatically detecting such information. Results are described for several speech domains and transcription conditions. Disfluencies are studied from a fluent and disfluent perspective, both on human subjective perceptions, and on the perspective of computerized classification results, based on hand annotated material.

Two distinct strategies are pointed for detecting disfluencyes, a contrastive strategy between the reparandum and the repair of fluency, manifested by pitch and energy increases at the onset of the repair, and a parallel prosodic strategy between this same areas, meaning, the repair mimics the tonal patterns of the reparandum Levelt (1983). The existence of an edit signal is confirmed, which is capable of denouncing an upcoming repair, producing cues such as: fragments; repetition patterns; glottalizations; co-articulatory gestures; voice quality attributes, such as jitter (perturbations in the pitch period) in the reparandum; differences pause durations from fluent boundaries; the occurrence of specific lexical items in the interregnum, and; pitch and energy increases in the repair. Several studies stress the importance of the boundaries of disfluency *vs.* fluency in the speech array to look for traces, and also the existence of contrastive characteristics between the disfluent region and the oncoming repair. For European Portuguese, it is reported a tendency towards repairing fluency using prosodic contrast making strategies, regardless of the disfluency genre, producing increases of both energy and pitch, which is confirmed in the results described in this dissertation. Filled pauses are pointed as the most distinct genre in therms of pitch increase, durational contours and, repetitions, in respect to energy rising patterns.

Several feature genres are reported as adequate for detecting disfluencies: prosodic, lexical, morphosyntactic informations and, semantic information. A multitude of features are reported for detecting disfluencies: pitch, pitch and energy slopes, the differences between the corresponding slopes, the tempo characteristics of the distinct regions and of the adjacent silent pauses, phrasing, the degree of cohesion between words, spectral characteristics, duration characteristics (word, pauses, syllables, phones), number of phones, modifications in segment durations, intonation, voice quality, vowel quality, coarticulation patterns, distance from a pause and, level of spoken word completion. Other reported features concern regular trends in disfluency relating to: sentence length, the presence of other disfluencies in the sentence, combinations of both these features both across and within speakers, the rate of the cut-off words and, the rate of editing phrases. Additionally, Shriberg and Stolcke (1996) suggests that speakers hesitate before less predictable words, pointing the utility of certain words transition probabilities as information source for discriminating disfluency. As for tempo patterns, it is reported that the region to repair is longer than the repair itself, and that there is a strong trend manifested in lengthy silences between these regions, and also the potential of pitch and energy slopes for disfluency type

differentiation, which seems to be independent of speaker gender. Some studies describe difficulties related to the classification of the distinct disfluent regions, pointing the interregnum as the easiest zone in terms of achieving a good detection performance, and the reparandum as the hardest. It is reported that the different disfluent regions are characterized by distinct prosodic properties. Concerning the repair region, the best linguistic features found are pitch and energy resets as well as the presence of a silent pause immediately before a repair.

As for filled pauses, the authors reported that elements that are considered disfluent are produced in a lower register, with descending contours that disrupt tonal scaling. Additionally it is reported that filled pauses tend to occur between the previous brake and the ongoing of the conversation, and are uttered at a tonal space in between adjacent prosodic constituents, that prosodic phrasing plays a crucial role in the task of classifying the perception of disfluency and, that contour shape is also important. The importance of the presence of a silent pause preceding the repetition, in the case of single filled pauses and prolongations, are also reported. The main features reported for detecting filled pauses are length, the distance from a pause, and pitch (fundamental frequency).

In what concerns classification methodologies several approaches are reported: maximum entropy models (CART, Logistic Regression), language model discrimination, Hidden Markov Model and, Conditional Random Fields, but CARTs generally achieve the best results.

3 Research Process and Data

This chapter describes the Corpus used in the experiments described in Chapter 4, the extracted feature set, as well as the corresponding extraction process. The following section describes the corpus used in the experiments performed in this dissertation. Section 3.2 describes a parser program, developed to extract the necessary representations for the target classes from the XML registries. Finally, Section 3.3 describes the set of features extracted from the XML references, using the parser described in Section 3.2.

3.1 Corpus

All work described in this dissertation is based on LECTRA Trancoso et al. (2008), a speech corpus of university lectures in European Portuguese, originally created for multimedia content production, and to support hearing-impaired students.

The speech signal is a rich source of information from which several information genres can be extracted. ASR systems use the speech signal to detect phones and transcribe them into an array of lowercase characters, grouped in words or word fragments, and separated by whitespace characters.

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 3.1: Properties of the Lectra Corpus.

Using the audio speech source, most ASR systems are able to calculate and store the time period of speech events such as words, syllables and, phones. Based on the computed intervals, and on the pitch, energy and, duration characteristics of the audio signal, a multitude of descriptors can be automatically computed: word confidence scores, syllable stress, speaker gender, background speech conditions (clean / noise), among others. However, the resulting automatic transcription contains a brute representation, that must be cleaned, ascertained in terms of consistency, and enriched with missing information by a reliable source such as linguists, in order to be legible, and for further processing tasks.

In the present work, the automatic transcripts were produced by the audio pre-processing and speech recognition modules of AUDIMUS (Meinedo et al., 2003; Meinedo, 2008). Among other encoding types, the resulting material is stored into a set of self contained XML files, to keep all the information required for further experiments in an interoperable format, possibly including for every word token information about, the time period corresponding to each word, confidence measures, background noise, speaker cluster, speaker gender and other metadata. Speech transcriptions are typically a combination of automatically acquired material, and orthographically transcribed information, performed by human annotators. The manual annotations account for structural meta-information such as disfluencies, disfluent types, disfluent regions, punctuation marks, capitalization, paralinguistic annotation and, further prosodic information, etc, which can either be seen as targets or features when performing classifications.

In the present work, the references concerning disfluencies were produced using AUDIMUS to obtain a forced alignment between the manual and automatic transcriptions, which are finally merged into single XML files. In this process, the calculated word boundaries are adjusted using prosodic features (pitch, energy, duration) and by applying post-processing rules. The transcription alignment task is not trivial due to the occurrence of recognition errors. To perform the alignment task the NIST SCLite tool[1] is used, followed by an automatic post-processing stage, for correcting possible SCLite errors and aligning special words which can be written / recognized differently. The framework used for producing the reference material is described in Batista et al. (2012).

Table 3.1 presents relevant information about the corpus concerning disfluencies, both for the training / development phase, which were performed in conjunction, and also for the data used in the test phase. The corpus contains records from seven 1-semester courses, where most of the classes are 60-90 minutes long, and consist of spontaneous speech. Due to a recent extension, the corpus contains about 32h of manual orthographic transcripts, which were split into 2 different subsets (training+development and test).

The work conducted in the scope of this dissertation is performed into 2 stages. In the first phase, our research aimed at compiling, based on the existing literature, a good feature set for detecting the disfluent region, it's inherent structures and, filled pauses (FPs). Literature for several languages was

[1] available from http://www.nist.gov/speech.

considered, since it is known that cross-language similarities exist in describing disfluent phenomena. The extracted set of features concern lexical and prosodic cues, but the vast majority relies on prosody, *vide* Section 3.3. A supervised learning approach was chosen based on several methods widely used and described in the literature, and also on the existing annotated data, and nature of the scope of feature data types. In a second phase, the focus shifts to applying different classification methods to the detection of FPs, a specific type of disfluency that is often mistaken with words in a language by automatic speech recognition systems.

The following subsection describes a parser program, developed to extract the features used in the experiments performed in this dissertation.

3.2 XML Parser

Since all transcribed data available from LECTRA Trancoso et al. (2008) corpus is stored in XML format, a parser program was written using the Java programing language, to extract the features and produce the input for Weka [2]. The XML registry contains both automatically acquired information arising from the ASR and manually transcribed information. Ah example of the material used for extraction is comprised in Figure 3.1, note that, empty lines were added simply for visualization purposes. The speech information must be extracted from a large amount of data stored in several generated arff files. A final version is then generated, comprising one general header and the remaining content is concatenated, producing the input for the Weka suite (Figure 3.2). In the prediction process, respecting the nature of the data is mandatory for engineering a faithful corresponding representation. Not doing so is guaranteed to compromise results of a real test, since several important information aspects are being ignored. For linguistic data, respecting the nature of the material concerns taking into account information that relates to higher structures than word or sentence, respecting the inherent relation of the word elements in the sequential pack of a conversation. In this view, dealing with sequential sentences may demand different extracting proceedings than when processing phrases, since phrases are always uttered in interconnection, and sentences may be separated by long silences, but still be semantically part of the same conversation pack. In the present work, all sentences are considered to be interconnected, and the calculation of boundary word variables is performed in a straight forward fashion.

As can be seen in the first row bellow the @*data* tag in Figure 3.2, some variable slots present a question mark instead of an outcome. This is foreseen by Weka as elements that can not be calculated, and are not treated as an outcome, but rather an empty value. Some feature calculations may be naturally impossible to calculate, such as the pitch of some phones, since in some cases a phone may be verbalized without pitch. Examples of question marks appear at the beginning and at the end of each

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

transcript segment. For instance, the first occurring question mark in the first line after the *@data* tag relates to the feature that quantifies the amount of time taken in the silence preceding the current word. Since the actual token represents the first word occurrence, a preceding silence cannot be calculated. Additionally, a question mark may be generated in cases where some feature cannot be calculated, such as in pitch slope comparisons, in cases where one of the corresponding values happens to be 0. This can happen when the vowel is cut off from a word while being spelled, as is example “de” (the), leaving only a consonant sound that is recognized with no pitch, since it is an unvoiced consonant. In the case of extracting features from sequential XML files, representative of different classes, for the first and last words of each XML file some features concerning the previous and following word cannot be calculated, since performing this calculation would violate the true continuity of the speech material. This issues demand the use of methodologies that can deal with unknown feature values, such as the ones tested in this work. Disfluency annotation is marked according to Shriberg (1994), and additional annotation was added containing information relative to the syntactic and prosodic context of the disfluencies, as well as those of all the silent pauses in the LECTRA.

The set of features extracted relates mostly to prosody, and can be consulted in the next section.

3.3 Feature Set

This section presents the extracted set of features, the feature scope focusses on the description of the current word, the surrounding silences, and first level neighboring words, from which information about pitch, segmentation, and energy, are extracted. From these sources most of the remaining descriptors are computed. A minority of cues accounts for lexical information related to word equality comparisons.

The proposed set of features encompasses distinct genres: numerical information (real or integer), binary, or sets of possible categorical outcomes. Some features such as $fragment$ (cw_isfrag in Figure 3.2) represent binary values, while other features comprise categorical possibilities, such as $pslope.w_0.w_{+1}$ (cw_fw_pslope in Figure 3.2). The categorical possibilities are defined inside brackets, surrounded by quotes and separated by commas. In the declaration of *cw_fw_pslopes* ($pslope.w_0.w_{+1}$) and *cw_fw_eslopes* ($eslope.w_0.w_{+1}$), the categorical values *’R’* and *’F’* represent the rise and fall pitch contours on the previous and following words, and the ’-’ character denotes situations where the pitch contour remains stable (plateau contours). The categorical possibilities of the features $silence.comp$ and $durationcomp.w_0.w_1$ (<, >, =) represent silence and duration comparisons between two consecutive units. These two features are consensually described in the literature as having a major impact in the identification of the different disfluent regions. The use of quotes is discarded in case of the predicted binary possibility yes / no ({*y, n*}). Note that, bellow the *@data* tag in Figure 3.2 the token “out” corresponds

to the *disfpos* categorical entry and denotes the end of the array of a particular word occurrence description. The feature set composition is described as follows: $filled.p$ (checks if the current word is a filled pause), $filled.p.w_{+1}$ (checks if the following word is a filled pause), $fragment$ (checks if the current word is a fragment), $conf.w_0$ (current word confidence score), $conf.w_{+1}$ (following word confidence score), $num.phones.w_0$ (number of phones of the current word), $num.phones.w_{+1}$ (number of phones present in the following word), $num.syl.w_0$ (number of syllable(s) in the current word), $num.syl.w_{+1}$ (number of syllable(s) present in the following word), $equality.w_0.w_{-1}$ (checks if the current word is exactly equal to the word said before), $equality.w_0.w_{+1}$ (checks if the current word is exactly equal to the next word), $dur.w_0$ (current word duration), $dur.w_{+1}$ (the duration of the word after the current word), $dur.comp.w_0.w_{+1}$ (a comparison between the duration of the current word and the duration of the next word), $dur.ratio.w_0.w_{+1}$ (the ratio between the duration of the current word and the duration of the following word), $e.min.w_0$ (current word minimum energy), $e.max.w_0$ (current word maximum energy), $e.med.w_0$ (current word median energy), $e.med.ratio.w_0.w_{+1}$ (the ratio between the median energy of the current word and the median energy of the following word), $e.dif.w_0.w_{-1}$ (the energy difference between the current word and the previous word), $e.dif.w_0.w_{+1}$ (the energy difference between the current word and the following word), $e.slope.w_0$ (current word energy slope), $e.slope.w_0.w_{+1}$ (compares energy slope values between the current word and the following word), $p.dif.w_0.w_{-1}$ (difference between the current word and the last in terms of pitch), $p.dif.w_0.w_{+1}$ (difference between the current word and the subsequent word in terms of pitch), $p.slope.w_0$ (the pitch slope shape of the current word), $p.slope.w_0.w_{+1}$ (compares the pitch slope shape of the current word and the one obtained by the next word), $p.med.ratio.w_0.w_{+1}$ (the ratio between the median pitch of the current word, and the median pitch of the following word), $b.sil.w_0$ (the duration of the silence before the current word), $b.sil.w_{+1}$ (the duration of the silence after the current word), $b.sil.comp.w_0.w_{+1}$ (a comparison between the silence before the current word and the one after), $b.sil.ratio.w_0.w_{+1}$ (a ratio between the silence duration before and after the current word). Initial tests have included the calculation of current word minimum pitch, current word maximum pitch and, current word median pitch, however the corresponding impact on classifications was not significant, and these where excluded from further experiments, for simplification. Note that the only lexical features are $equality.w_0.w_{+1}$ and, $equality.w_0.w_{-1}$. The resulting extraction file is used as input in Weka[3], where the data can be visualized, and algorithm configuration and training performed. Figure 3.2 shows the resulting arff file.[4] The empty line separating the 1º and 2º data rows, was added for visualization purposes.

The use of filled pauses (FPs) and fragments (FRG) as features, is motivated by the fact that these are the most common disfluencies, known to largely populate the disfluent regions. Filled pauses are more abundant in the interregnum, while FRGs tendentiously occur before this zone. Therefore, both these features are expected to have a big impact on classifications, which makes them a priority for build-

[3]Weka version 3-6-8. http://www.cs.waikato.ac.nz/ml/weka
[4]Arff file syntax - http://www.cs.waikato.ac.nz/ml/weka/arff.html

```
(...)

<TranscriptSegment>

<TranscriptGUID>1</TranscriptGUID>
<AudioType conf="1.000000" end="2560" start="2536">Clean</AudioType>
<Time end="2560" reasons="" sns_conf="1.000000" start="2536"/>
<Speaker gender="U" gender_conf="1.000000" id="0" id_conf="1.000000" known="T" name="
    Unknown"/>
<SpeakerLanguage native="T">PT</SpeakerLanguage>

<TranscriptWordList ph_avg="5.0" ph_duration="10" phones="2" syl_avg="10.0" syl_duration="
    10" syls="1">

 <Word conf="0.994262" eavg="46.5" emax="52.7" emed="49.4" emin="35.0" end="2552" eslope="
     2.0" focus="F1" name="bom" pavg="140.9" phseq="_bo~+" pmax="152.7" pmed="141.0" pmin
     ="128.8" pos="A." ps_med="-5.9" pslope="-5.9" punct="." start="2543">

 <syl dur="10" eavg="46.5" emax="52.7" emed="49.4" emin="35.0" eslope="2.0" pavg="140.9"
     pmax="152.7" pmed="141.0" pmin="128.8" ps_med="-5.9" pslope="-5.9" start="2543"
     stress="y">

 <ph dur="4" eavg="39.5" emax="43.7" emed="39.6" emin="35.0" eslope="2.7" name="b" pavg="
     0.0" pmax="0.0" pmed="0" pmin="0.0" ps_med="0.0" pslope="0.0" start="2543"/>

 <ph dur="6" eavg="51.1" emax="52.7" emed="51.7" emin="47.5" eslope="0.7" name="o~" pavg="
     140.9" pmax="152.7" pmed="141.0" pmin="128.8" ps_med="-5.9" pslope="-5.9" start="
     2547"/>

</syl>
</Word>

</TranscriptWordList>

</TranscriptSegment>

(...)
```

Figure 3.1: XML excerpt example.

ing a disfluency detection module. Comparing results of including and excluding these elements allows both to understand the corresponding weight on classifications, and also to evaluate the expressiveness of the proposed acoustic feature set on the classification tasks. 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. Note that, not knowing whether the current element is a FRG or a FP may have a strong impact in the results.

```
@relation disf
@attribute cw_isfrag {y, n}
@attribute cw_isfp {y, n}
@attribute cw_pslope real
@attribute cw_emin real
@attribute cw_emax real
@attribute cw_emed real
@attribute cw_eslope real
@attribute cw_conf real
@attribute cw_dur real
@attribute cw_phones numeric
@attribute cw_syls numeric
@attribute cw_bsil real
@attribute fw_isfp {y, n}
@attribute fw_pslope real
@attribute fw_eslope real
@attribute fw_conf real
@attribute fw_dur real
@attribute fw_phones numeric
@attribute fw_syls numeric
@attribute fw_bsil real
@attribute cw_fw_sil_comp {"<", ">", "="}
@attribute cw_fw_dur_comp {"<", ">", "="}
@attribute cw_fw_equals {y, n}
@attribute cw_fw_pslopes {"RR","R-","RF","-R","--","-F","FR","F-","FF"}
@attribute cw_fw_eslopes {"RR","R-","RF","-R","--","-F","FR","F-","FF"}
@attribute cw_fw_pdiff real
@attribute cw_fw_ediff real
@attribute pw_cw_pdiff real
@attribute pw_cw_ediff real
@attribute pw_cw_equals {y, n}
@attribute cw_dur_fw_dur_ratio real
@attribute cw_bsil_fw_bsil_ratio real
@attribute cw_pmed_fw_pmed_ratio real
@attribute cw_emed_fw_emed_ratio real
@attribute disf {y, n}
@attribute ip {y, n}
@attribute interregnum {y, n}
@attribute repair {y, n}
@attribute disfpos {"ip", "int", "disf", "repair", "out"}
@data
n,n,-5.9,35.0,52.7,49.4,2.0,0.994262,9,2,1,?,n,3.3,1.7,0.988117,8,5,1,34,?,">",n,"FR","RR
    ",6.47,-5.7,?,?,?,0.53,?,0.41,0.53,n,n,n,n,"out"

n,n,3.3,26.4,45.4,43.7,1.7,0.988117,8,5,1,34,n,-4.3,-0.2,0.990069,14,5,2,1,">","<",n,"RF
    ","R-",0.87,0.4,6.47,-5.7,n,0.36,0.97,0.49,0.5,n,n,n,n,"out"
(...)
```

Figure 3.2: Arff excerpt example.

4 Experiments

This chapter describes our experiments concerning the automatic detection of disfluencies and their structural elements, and also initial experiments on the detection of *filled pauses*. Experiments are performed using Weka[1], an open source collection of machine learning algorithms and tools for data preprocessing and visualization. Several approaches are tested, namely: Naïve Bayes, Logistic Regression, Multilayer Perceptron (MP), J48 or Classification and Regression Trees (CART). Algorithm default configuration parameters where used, as contained in Weka (3-6-8).

The reminder of this chapter is structured as follows: Section 4.1 starts with the presentation of all the considered metrics. Section 4.2 reports on the first binary experiments aiming at automatically identifying which words belong to a disfluent sequence. Section 4.3 describes the results of the binary experiments in detail, and also the results a multi-class 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 multi-class classification, details relative to distinct disfluent zone classification performance will be presented. Finally, Section 4.4 describes binary classification experiments aimed at automatically identifying *filled pauses*, using a set of acoustic features. Disfluencies are classified according to Shriberg (1994), being the most frequent categories: repetitions, FPs, prolongations, deletions, substitutions, inserts, fragments (FRGs) and, editing expressions.

4.1 Evaluation Metrics

$$\textbf{Accuracy} = \frac{TP+TN}{N} \quad (1)$$

$$\textbf{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\textbf{Recall} = \textbf{Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

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

	Condition Positive	Condition negative	Total
Test outcome Positive	**TP**	**FP**	TP + FP
Test outcome negative	**FN**	**TN**	FN + TN
Total	TP + FN	FP + TN	**N**

Table 4.1: Matrix Example.

$$\textbf{fpRate} = \textbf{Fallout} = \frac{FP}{FP+TN} \quad (4)$$

$$\textbf{F1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

$$\textbf{Chance} = (\frac{TP+FP}{N}) \cdot (\frac{TP+FN}{N}) + (\frac{FN+TN}{N}) \cdot (\frac{FP+TN}{N}) \quad (6)$$

$$\textbf{KAPPA} = \frac{Recall - Chance}{1 - Chance} \quad (7)$$

$$\textbf{SER} = \frac{FP+FN}{TP+FN} \quad (8)$$

$$\textbf{tnRate} = \textbf{Specificity} = \frac{TN}{FP+TN} \quad (9)$$

This subsection describes the statistical measures of inter-rater agreement used in this chapter to evaluate results. Table 4.1 comprises a graphic example of a classification matrix, a standard tool for statistical model evaluation. In classification experiments such as the ones performed in the present chapter, a model (Test) sorts all cases present in a population (word tokens in this case) into categories. The results are matched against a reference (Condition) resulting in 4 distinct classification situations, as exemplified in Table 4.1: true positives (corrections / correctly classified slot occurrences, *TP*), false positives (insertions / false acceptances*, FP*), false negatives (deletions / false rejections, *FN*), true negatives (substitutions / correctly classified non slot occurrences*, TN*). Note that in Table 4.1, *N* referees to the total amount of elements in the population (word tokens). In multi-class classifications a model is given a multitude of slots to rate, resulting in an expanded version of the matrix present in Table 4.1.

Standard evaluation metrics were applied, encompassing: accuracy (equation 1), precision (2), recall (3), f-measure (5) and, Slot Error Rate (8). The metrics used in the present work are based on slots, which correspond to the elements that are the target for classification. 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 in an occasion where a disfluency is in fact present, represented as *TP* in Table 4.1. Since the aim is on classifying slots, only a minority of metrics involve the use of true negatives, which are usually present in much larger proportion in comparison to slot occurrences. Note that, obtaining too many substitutions (TN) is in fact undesirable, since this may denounce an overfitting case, which means the method has adapted too much to the prevalent class, resulting in a naive tendency towards performing slot classifications. The intrinsic asymmetry of *TP vs. TN* present in the experiments performed

throughout this chapter, makes it much more difficult to get a good precision than a good specificity, while avoiding sensitivity (Recall) alterations. Since there are much more irrelevant occurrences than relevant in the corpus, there are also more occasions for deletions (FN) than for corrections (TP), and the insertion occurrences can overcome the true positive amount even if the classifier has impressive accuracy on a balanced size class test set.

Accuracy can be described as the proportion of the true results (true positives and true negatives) resultant from classifications, against the total population (N), being 100% when the measured values are exactly the same as the given values. It is a known fact that a high accuracy is not synonym for predictive power, and that it is possible a corresponding lower accuracy record may in fact reflect better performance. For this reason other metrics such as precision and recall are preferred.

Precision and recall are important base metrics that are concisely interconnected, presenting an explicit trade-of that is a prevalent element in classification results, different trade-off points between these metrics are appropriate in different situations. Both these metrics are slot oriented, meaning that *TN* is not used for calculations. Precision, also called positive predictive value, is the proportion of true positives against all positive results (TP + FP), which may also be seen as the probability that a positive classification is relevant (a correction), accounting for insertion errors. A hight precision means that the algorithm returned more relevant than irrelevant results. Recall on the other hand is the probability that a relevant classification is performed, accounting for the deletion errors. A hight recall means that the algorithm returned most of the relevant material. In choosing between approaches, the best performance depends on the specifics (application goals), and a trade off between these metrics dictates the final choice, depending on the preferred: positive examples or on positive predictions. A good precision is preferred when reliable classifications are important, such as in the case of fraud detection or natural catastrophe prediction tasks, while recall is better for capturing higher amounts of slot occurrences, being preferred for research tasks such as searching a hard disk for information.

F-measure is the standard way of combining these metrics, being defined as the weighed harmonic mean between precision and recall. In situations where the least prevalent class is more important, F-measure may be more appropriate than precision and recall, especially in cases with very skewed class imbalance. Calculating f-measure generates a value that is closer to the minimum of the two numbers than either the corresponding arithmetic and geometric means, making f-measure a very conservative measure / average. Makhoul et al. (1999) reports “this measure (f-measure) implicitly discounts the overall error rate, making the systems look like they are much better than they really are”. Recall, precision and f-measure have been target of several critiques: they ignore performance in correctly handling negative examples (TN), they propagate the underlying marginal prevalences and biases and, they fail to take into account the chance level performance. For the presented reasons, the preferred performance metric for performance evaluation is the SER.

The SER measure accounts for combining the different types of error directly (FP, FN), without having to resort to precision and recall as preliminary measures. This metric can be described as the division of the total number of non slot decisions (insertions, deletions), by the total number of slots (corrections and deletions). Note that precision, recall and f-measure produce values ranging from 0 and 1, while SER produces values higher than 1 whenever the number of errors exceeds the number of slots in the reference (condition).

Apart from the above mentioned performance metrics, two other additional metrics are often reported in this document. *Kappa* represents a way of debiasing and renormalizing accuracy, providing a notion of whether a classifier is doing better than chance. In the KAPPA equation (7), *c* represents the chance level decision or the hypothetical probability of chance agreement, which is displayed in Equation (6), being produced by using the observed data to calculate the probabilities of the observer randomly guessing each event.

The following two metrics, although not directly used in this work, are involved in the calculation of the ROC area. The false positive rate, also known as false alarm rate, represents the probability of falsely rejecting the null hypothesis, it can be seen as the expectancy of the false positive ratio. The true positives rate, also known as corrections rate, measures the proportion of actual positives, being complementary to the false negative rate. Note the true positives rate is also known as sensitivity (recall), while specificity (true negative rate) denotes the proportion of negatives which are correctly identified. Specificity can be seen as precision from the perspective of a TN rather than from a TP. Receiver Operating Characteristic (ROC) is a metric based on performance curves that can also be used for more adequate analysis (Liu and Shriberg, 2007). This metric consists of plotting the false alarm rate on the horizontal axis, while the correct detection rate is plotted on vertical, and calculating the area bellow the resulting line (Fawcett, 2006). This performance metric provides a notion of the relation between the amount of risk taken and the amount of correct classifications. The results related to this metric account for the area bellow the curve, a value of 0.5% represents chance agreement, and higher values represent improvements above chance. Methods based on trees do not provide probabilities over the classes and for that reason the corresponding ROC area cannot be fairly computed.

4.2 Detecting Elements that Belong to Disfluent Sequences

This section presents a high level performance analysis on results for detecting elements belonging to disfluent sequences, using forced alignment data and including filled pauses (FPs) and fragments (FRGs) as features. The classification judgements are always performed at the word level, resulting in

	Method	Time Train	Time Test	Accuracy	Kappa
IP	ZeroR	0.1	3.5	97.06	0.000
	Simple CART	3665.4	2.2	97.83	0.489
	J48	3810.9	1.9	97.79	0.485
	Logistic Regression	47.4	2.3	97.69	0.464
	Multilayer Perceptron	6268.7	15.1	97.59	0.453
	Naïve Bayes	741.0	4.1	92.05	0.214
Int	ZeroR	0.1	34.9	98.50	0.000
	Simple CART	708.7	2.1	99.95	0.982
	J48	552.0	2.7	99.95	0.982
	Logistic Regression	67.4	41.2	99.94	0.980
	Multilayer Perceptron	9176.3	9.6	99.92	0.974
	Naïve Bayes	723.5	4.4	99.24	0.779
Repair	ZeroR	0.1	3.0	97.06	0.000
	Simple CART	2688.9	2.1	97.27	0.203
	J48	2364.6	2.4	97.30	0.207
	Logistic Regression	42.5	2.5	97.21	0.204
	Multilayer Perceptron	6348.3	9.9	97.25	0.187
	Naïve Bayes	553.0	5.4	88.54	0.133
Disf	ZeroR	0.1	3.2	91.67	0.000
	Simple CART	3412.4	2.0	94.44	0.502
	J48	3818.5	1.9	94.37	0.505
	Logistic Regression	40.5	3.1	94.40	0.503
	Multilayer Perceptron	8473.2	9.2	93.93	0.489
	Naïve Bayes	551.9	5.6	89.84	0.362
DisfPos	ZeroR	0.1	3.4	88.73	0.000
	Simple CART	6148.8	1.8	91.55	0.420
	J48	4602.1	1.9	91.39	0.414
	Logistic Regression	1391.1	3.9	91.36	0.416
	Multilayer Perceptron	10209.7	12.3	91.40	0.414
	Naïve Bayes	574.7	7.4	76.48	0.223

Table 4.2: High level performance analysis for detecting elements belonging to disfluent sequences using alignment data and including FP and FRG as features.

the classification of the current word as one of the possible classification outcomes. The performance results presented in Table 4.2 are obtained using the forced aligned version of the data, the following section comprises results for corresponding automatically recognized material. The first column of this table presents the regions targeted for classification.

All disfluent related regions are targeted in separate binary classifications, namely: the interruption point (IP), the interregnum region (Int), and the repair (Repair). *Disf* represents a binary classification, aimed at classifying words that rely inside the disfluent event boundary. Finally*, DisfPos* represents a multi-class classification experiment, in which all the disfluency related regions targeted in the binary experiments are treated as possible classification slots, plus a slot for fluent elements other than those found inside the repair. In this work, the previous word to the *IP* event is used to discriminate this region, note that the *IP* is not definable as a word, but by marginal properties of the words that surround the gap, and the properties of the gap itself. The second column of Table 4.2 presents the applied methods per

classification zone. The subsequent columns represent measurement indicators. *Time Train* presents the amount of time consumed in model building, and *Time Test* represents the time required for applying the model. The percentage of correctly classified Instances (Accuracy) considers all the elements that are being classified and not only slots. *ZeroR* represents the baseline obtained by forwarding all classifications towards the predominant classification category, by ignoring all predictors, thus providing a baseline for performance comparison and evaluation.

Regarding the binary classifications, the results displayed in Table 4.2 reveal the *IP* and *Disf* regions are the most extensive in terms of training time consumption. As for *Accuracy* and *Kappa*, the most affordable area concerns the interregnum, followed by the interruption point, for which a similar performance as the one seen for the repair region is achieved, followed by the classification of the whole disfluent region. Although based on these metrics the performance of the multi-class experiments seems to be lower than the ones achieved for the disfluent regions individually, an analysis performed in the following section on the same results reveals that the performance achieved for each individual region is superior to accounting for the region individually in a binary classification. For the binary classification of distinct in-disfluency areas, sometimes the repair obtains similar accuracy to the one recorded in the classification of interruption point, although slightly lower. However the latter presents much higher *Kappa* values, suggesting the adopted set of features provides a better representation for this zone. Classifying the interregnum generally results in close to perfect performance, a fact that is associated with the inclusion of FPs as features, which strongly characterize this zone. Detecting the whole disfluent region (Disf) culminates in the worst classification results within the binary classifications.

For classifying all regions simultaneously (DisfPos) we assist to a general increase in time required for model building, and better classification results in comparison to the binary classification of any other zone individually. It seems that considering all zones for classification accounts for increased disambiguations. In general, the required time for model building increases along with the increase in slot number, especially for Multilayer Perceptron, while test time does not suffer considerable variations. Multilayer Perceptron is the most costly approach in terms of training time, achieving similar results to J48, while consuming more than twice the time taken by the latter approach. Although CARTs achieve the best results among the multi-class classifications, Logistic Regression presents a kappa value greater than J48 and Multilayer perceptron. This value is identical to the best result achieved with CARTs, at cost of a widely reduced time interval than any other approach.

In comparison to the interregnum region, all areas present reduced *Kappa* records. As regards the accuracy metric, the *Disf* region is the hardest to reliably detect, all the remaining classification targets obtain better Accuracy. The interregnum region contrasts with this trend, displaying a smaller difference. The results presented show that Multilayer Perceptron consistently requires far more training time than the remaining approaches, frequently resulting in mediocre performance when compared to

the remaining approaches. In fact, Logistic Regression achieves similar results while performing much briefly in the training phase. However, based on these metrics this inference is not conclusive, due to both the closeness of the achieved results and the potentially deceiving nature of both *Kappa* and Accuracy parameters. Naïve Bayes clearly presents the less suited approach, resulting in much lower *Kappa* values. CARTs consistently provide best performance. Despite generally performing slightly lower in comparison to other state of the art methods, Logistic Regression may present considerable advantage in situations where large amounts of data are considered, since this approach consistently produces timely and acceptable results. Both decision trees consistently perform better than the remaining methods while requiring considerably less time for training than Multilayer Perceptron, its nearest competitor in therms of performance.

The next section holds a more detailed and performance oriented analysis towards the discrimination of the distinct regions of a disfluency.

4.3 Detecting and Distinguishing Elements Between the Disfluent Regions

This section presents results for the binary classifications of the disfluent regions, including the disfluent region as a whole, an overview of binary performances, a multi-class experiment, a summary of results achieved in these experiments and, a final subsection comprising feature impact analysis.

The following subsections explore results of classification experiments described in the previous section in a more performance oriented fashion. Comparable results are obtained for the same tasks using recognized data. The reminder of this section is organized as follows: Subsections 4.3.1 to 4.3.6, perform individual result analyses using 4 distinct setup combinations, varying in the use of either, forced alignment material, or raw ASR recognition data, and the inclusion or not of filled pauses (FPs) and FRGs as features; Subsection 4.3.1 refers to the interruption point detection task; Subsection 4.3.2 refers to the interregnum; Subsection 4.3.3 refers to the repair region; Subsection 4.3.4 refers to detecting the disfluent region in it's totality; Subsection 4.3.5 provides an overview on the binary classification results, which are performed by summing the amounts of, corrections, and insertions, from the results of all the binary experiments, and using the data to compute the relevant metrics described in Section 4.1; Subsection 4.3.6 is somewhat different, since it additionally contains the detailed analysis of the best results achieved in the multi-class experiment, achieved with CART for the alignment data while excluding FPs and FRGs as features, and finally the corresponding result matrix analysis. These experiments target the disfluent regions, including a slot for the reparandum, and a slot for elements located outside of disfluent sequences; Subsection 4.3.7 presents a summary and conclusions; Finally, Subsection 4.3.8

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	762	71	1899	81.5	27.6	41.2	75.0	
	J48	749	96	1891	77.4	27.1	40.2	76.4	
	LR	763	118	1883	76.5	27.6	40.6	76.7	0.79
	MP	753	99	1891	77.5	27.3	40.3	76.3	0.78
	NB	980	3983	1317	18.1	35.5	23.9	208.7	0.73
Recognition	CART	17	6	684	73.9	2.4	4.7	98.4	
	J48	41	83	660	33.1	5.8	9.9	106.0	
	LR	21	22	680	48.8	3.0	5.6	100.1	0.78
	MP	72	127	629	36.2	10.3	16.0	107.8	0.72
	NB	128	795	573	13.9	18.3	15.8	195.1	0.72

Table 4.3: Detailed performance analysis on predicting the interruption point obtained while including FP and FRG as features.

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	85	34	635	71.4	11.8	203	92.9	0.60
	J48	71	51	649	58.2	9.9	16.9	97.2	0.80
	LR	75	63	645	54.3	10.4	17.5	98.3	0.79
	MP	221	1672	499	11.7	30.7	16.9	301.5	0.77
	NB	73	38	647	65.8	10.1	17.6	95.1	0.57
Recognition	CART	16	9	685	64.0	2.3	4.4	99.0	0.53
	J48	16	30	685	34.8	2.3	4.3	102.0	0.52
	LR	3	13	698	18.8	0.4	0.8	101.4	0.75
	MP	23	31	678	42.6	3.3	6.1	101.1	0.74
	NB	102	865	599	10.5	14.6	12.2	208.8	0.67

Table 4.4: Detailed performance analysis on predicting the interruption point obtained while excluding FP and FRG as features.

presents a feature impact analysis of the top-20 cues for the experiments performed on forced alignment material, while including or excluding FPs and FRGs, focussing on both feature weight and potential for distinguishing between zones.

4.3.1 Interruption Point Detection

Table 4.3 comprises results for detecting the interruption point region, regarding: the *Cor*, correct classifications; *Del*, marked in the reference but not correctly classified; and Insert slots, not marked in the reference. The values presented for precision, *Prec*, recall, *Rec*, f-measure, *F,* and Slot Error Rate, *SER* represent percentages. CART and J48 are not probabilistic classifiers, therefore the ROC curve area can not be fairly computed.

The quality of the alignment results can be seen in linear fashion as the quality of method performance decreases from the top to the bottom of Table 4.3. Regarding this data, CART stands out as the best option for detecting the interruption point region, performing better in almost all metrics. In fact, the same applies for the CART's results achieved including filled pauses and fragments, and using recognition data, but in this case the performance difference in comparison to the remaining methods is quite larger.

For the alignment data experiments, performed including filled pauses and fragments as features, CART consistently performs better than J48, obtaining considerably better precision (the best record), and good recall. This suggests that, given reliable data CART tends to overcome J48 in therms of assertiveness. Multilayer Perceptron risks even more than Logistic Regression, while achieving residual gain in terms of correct slot classifications, in conjunction with considerably higher insertion rates and higher SER ratings. Naïve Bayes takes by far the highest risk amount among the tested approaches, presenting highly dilated error rates for both insertions and deletions, failing even for the detection of the most common occurrences. Multilayer Perceptron risks even more than Logistic Regression, resulting in residual gain for correct slot classifications, in line with a considerably higher insertion level, and worse SER performance. Under this conditions, the best approach for obtaining increased precision is CART, while Logistic Regression accounts for better recall.

For the alignment data related experiments, excluding filled pauses (FPs) and fragments (FRGs) as features, mostly affects the recall metric, having a blunt impact on overall performance. This trend prevails throughout the results achieved using recognition data, while excluding these features. For the alignment experiments the removal of these two features results in recall losses prowling on average 10 units when accounting for Multilayer Perceptron, and 13 otherwise, given that the corresponding recall score is very discrepant in comparison to the remaining methods. The precision metric is also severely affected with the removal of these features, losing an average of 14 units when accounting for all tested methodologies. Concerning precision and recall the smallest loss belongs to CART, contrasting with Logistic Regression, holder of the sharpest deterioration. In general, we assist to severe decreases for corrections and increased deletions, showing the expressiveness of FPs and FRGs for discriminating this region. Given these conditions, CART presents the best results, providing considerably better precision and recall than the remaining methods, except for Naïve Bayes that generally presents high recall values at the expense of a pronounced precision loss.

For the recognition experiments, achieved including FPs and FRGs the results are not linear and the best choice is debatable. CART stands out as the best choice in terms of assertiveness, featuring quite larger precision values than the remaining methods. Nonetheless, the corresponding recall result suggest that this method tends to classify correct slot occurrences only in the presence of a high level of certainty, resulting in lower SER and correction records in comparison to the remaining approaches.

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	366	12	1	96.8	99.7	98.3	3.5	
	J48	366	12	1	96.8	99.7	98.3	3.5	
	LR	364	12	3	96.8	99.2	98.0	4.1	1.00
	MP	359	11	8	97.0	97.8	97.4	5.2	1.00
	NB	337	157	30	68.2	91.8	78.3	51.0	0.99
Recognition	CART	199	60	48	76.8	80.6	78.7	43.7	
	J48	200	63	47	76.0	81.0	78.4	44.5	
	LR	210	54	37	79.5	85.0	82.2	36.8	1.00
	MP	204	55	43	78.8	82.6	80.6	39.7	1.00
	NB	223	105	24	68.0	90.3	77.6	52.2	0.99

Table 4.5: Detailed performance analysis on predicting the interregnum obtained while including FP and FRG as features.

This could represent an advantage depending on the application targets. In comparison to CART, J48 risks markedly more when facing data of this nature, resulting in a much higher level of corrections / insertions, followed by fewer deletion levels. Although Logistic Regression offers the most balanced option, the corresponding corrections amount is not much higher in comparison with CART, contrasting with a markedly lower insertion rate in the latter case, and also a slightly higher number of deletions. This seems to point CART as a slightly superior approach for these conditions. Multilayer Perceptron seems to be the best option for hedging data. Nonetheless, this approach also presents a significantly higher error level than the other approaches (except Naïve Bayes), and much lower ROC area in comparison to Logistic Regression.

For recognition the removal of FPs and FRGs results in a large loss of precision for Logistic Regression, matched by an almost negligible recall loss. Interestingly, Multilayer Perceptron gains in precision but loses a slightly higher recall amount, outperforming J48. CART's achieve the best results obtaining a much higher accuracy than Multilayer Perceptron in conjunction with slightly lower recall. However Multilayer Perceptron achieves a considerable larger amount of corrections and insertions in comparison to CART, together with a smaller number of deletions, suggesting it is a valuable choice if data coverage is preferred.

In general, results obtained under these conditions are quite weak, as can be seen via baseline comparison present in Table 4.2. CART is the only method presenting results above this threshold, but still the improvement does not exceed half a percentage point.

4.3.2 Interregnum Detection

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	155	69	212	69.2	42.2	52.5	76.6	
	J48	195	107	172	64.6	53.1	58.3	76.0	
	LR	135	56	232	70.7	36.8	48.4	78.5	0.97
	MP	178	92	189	65.9	48.5	55.9	76.6	0.97
	NB	253	988	114	20.4	68.9	31.5	300.3	0.94
Recognition	CART	94	48	153	66.2	38.1	48.3	81.4	
	J48	120	65	127	64.9	48.6	55.6	77.7	
	LR	99	40	148	71.2	40.1	51.3	76.1	0.99
	MP	131	103	116	56.0	53.0	54.5	88.7	0.98
	NB	186	519	61	26.4	75.3	39.1	234.8	0.96

Table 4.6: Detailed performance analysis on predicting the interregnum obtained while excluding FP and FRG as features.

This subsection comprises results and analysis for the detection of the interregnum region. As expected the interregnum results are very satisfactory, mostly due to the presence of filled pauses (FPs) as features, as it can be seen in Tables 4.5 and 4.6. Note that this zone is largely populated by these elements. Roughly all alignment results present in Table 4.5 are close to perfect, except the ones registered for Naïve Bayes. It seems this method has a consistent tendency towards slot classification, resulting in widely dilated insertion and error rates. This trend prevails in the recognition data results, except in this case parameter values do not vary as much, which does not occur for the remaining methods. This may indicate hight insensitivity to this type of data, revealing inappropriateness for this kind of task. Since both best methods performed identically (CART and J48), the tiebreaker is the time consumption for model construction. Based on this premise, J48 is best suited approach, accounting for less insertions and deletions, and taking considerably less time for model construction. To account for fast model building, note that Logistic Regression presents suitable results while consuming solely approximately 70 seconds, achieving better SER than Multilayer Perceptron while consuming far less time in the building phase.

Removing filled pauses (FPs) and fragments (FRGs) from the alignment data experiments results in drastic performance losses at all levels, producing a severe impact on both precision and recall, in all cases. All metrics suffer a sharp increase, emphasizing the importance of FPs and FRGs for the characterization of this area. Note a sharp increase in all cases for the occurrences of insertions and deletions. J48 clearly seems the best option, holding a combination of the best: recall, f-measure, and SER results. The best precision record is achieved using CART, but at cost of a considerable recall loss. The following best approach in this case is Multilayer Perceptron costing far more time for training than J48.

The recognition outcomes comprised in Table 4.5, show remarkable results for Logistic Regression, revealing an advantaged performance over other methods in all metrics. Interestingly Multilayer Perceptron performs better than CART and J48, countering the trend experienced in most classifications.

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	89	38	631	70.1	12.4	21.0	92.9	
	J48	90	31	630	74.4	12.5	21.4	91.8	
	LR	92	55	628	62.6	12.8	21.2	94.9	0.80
	MP	81	34	639	70.4	11.3	19.4	93.5	0.74
	NB	293	2382	427	11.0	40.7	17.3	390.1	0.73
Recognition	CART	13	8	686	61.9	1.9	3.6	99.3	
	J48	22	14	677	61.1	3.1	6.0	98.9	
	LR	5	7	694	41.7	0.7	1.4	100.3	0.71
	MP	50	225	649	18.2	7.2	10.3	125.0	0.66
	NB	84	582	615	12.6	12.0	12.3	171.2	0.67

Table 4.7: Detailed performance analysis on predicting the repair obtained while including FP and FRG as features.

Comparing CART and J48 reveals the former performs slightly better.

For recognition the removal of FPs and FRGs does not impact results as heavily as for alignment, as can be seen in Table 4.6. However the results remain consistently below those seen for the alignment data experiments. Removing these features leads to a general correction occurrences reduction to half, together with a marked deletion increase. For insertions, these values do not suffer large changes. Multilayer Perceptron seems to obtain the most balanced results, however, the corresponding f-measure and SER value remain bellow the one obtained by J48. Although Multilayer Perceptron achieves a higher correction record than J48, the latter obtains better precision, and a significantly lower insertion value. Apart from Logistic Regression, all approaches suffered considerable training time increases, in consequence to the removal of FPs and FRGs. Although Multilayer Perceptron achieves more corrections, J48 obtains better precision and a significantly lower insertions amount. The best approach is J48, presenting better recall then CART, more corrections and, less SER.

4.3.3 Repair Detection

This subsection, describes results for the detection of the repair region.

Table 4.7 shows that J48 stands as the best solution for the alignment data, while including filled pauses and fragments, obtaining a sharper higher precision amount than CART, which seems to be the second best choice, and also better recall and SER. Logistic Regression achieves a non-expressive highest correction rate, but the corresponding recall relies well bellow the one achieved by J48. Despite performing inferiorly, the corresponding timely results may present sufficient motivation for choosing

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	81	36	639	69.2	11.3	19.4	93.8	
	J48	80	31	640	72.1	11.1	19.3	93.2	
	LR	88	52	632	62.9	12.2	20.5	95.0	0.79
	MP	85	40	635	68.0	11.8	20.1	93.8	0.74
	NB	292	2402	428	10.8	40.6	17.1	393.1	0.77
Recognition	CART	11	7	688	61.1	1.6	3.1	99.4	
	J48	23	19	676	54.8	3.3	6.2	99.4	
	LR	5	7	694	41.7	0.7	1.4	100.3	0.71
	MP	38	89	661	29.9	5.4	9.2	107.3	0.67
	NB	83	580	616	12.5	11.9	12.2	171.1	0.67

Table 4.8: Detailed performance analysis on predicting the repair while excluding FP and FRG as features.

this approach, since Logistic Regression performed 64 times faster in the model building phase. J48 consistently outperforms Multilayer Perceptron in all metrics, while consuming less time in model training. Naïve Bayes risks too much, resulting in large insertion amounts, contrasting with the corresponding corrections and deletions, resulting in poor classification results. The results for alignment comprised in Table 4.8 reveal that, filled pauses (FPs) and fragments (FRGs) don't seem to have a blunt impact on results.

The results of alignment comprised in Table 4.8 show that the removal of FPs and FRGs negatively impacts overall SER records. Multilayer Perceptron and J48 perform similarly, obtaining an identical SER value. Multilayer Perceptron accounts for slightly better recall, while J48 obtains a better precision. The f-measure value obtained by Multilayer Perceptron points this approach performs slightly better than CART. J48 outperforms CART under these conditions, obtaining a combination of: identical corrections amount, lower SER and, the best precision record among all other methods for these conditions. Logistic Regression presents the best option for hedging data, presenting a higher ROC area that Multilayer Perceptron, and also an increased corrections amount. However, Multilayer Perceptron obtains a better SER record and also a fairly similar amount of corrections, suggesting this method presents a more balanced approach.

For recognition, including FPs and FRGs slightly affect results. J48 offers the best option, although none of the decision trees achieve sparkling results. CART provides slightly better precision, but this improvement is dwarfed by the amount recorded in the recall metric. Logistic Regression obtains considerably lower overall results than both decision trees, featuring a marked deletion rate and very low correct classifications. Multilayer Perceptron is the main surprise obtaining a very high correction amount, followed by a marked SER rate. However, CARTs achieve very similar results to Multilayer Perceptron, while consuming less than half the time for training.

For recognition, excluding FPs and FRGs results in slight recall losses, while the precision decays are somewhat more pronounced. The same trend seen for Multilayer Perceptron results, present in

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	754	73	1289	91.2	36.9	52.5	66.7	
	J48	778	115	1265	87.1	38.1	53.0	67.5	
	LR	765	95	1278	89.0	37.4	52.7	67.2	0.80
	MP	799	244	1244	76.6	39.1	51.8	72.8	0.78
	NB	891	1339	1152	40.0	43.6	41.7	121.9	0.77
Recognition	CART	284	115	1783	71.2	13.7	23.0	91.8	
	J48	301	184	1766	62.1	14.6	23.6	94.3	
	LR	257	102	1810	71.6	12.4	21.2	92.5	0.70
	MP	306	136	1761	69.2	14.8	24.4	91.8	0.69
	NB	461	825	1606	35.8	22.3	27.5	117.6	0.68

Table 4.9: Detailed performance analysis on predicting the disfluent region as a whole while including FP and FRG as features.

Table 4.7 for recognition data, is observed, in respect to obtaining much better correction values than the remaining approaches, except in this case the negative SER impact is considerably less severe, but still well above the record obtained by Logistic Regression. The results of Logistic Regression remained intact except for a very slight improvement in the ROC metric, suggesting that the removal FPs and FRGs impacts positively the classification performance. For the decision trees the overall impact of removing these features is consistently negative, although not very pronounced. Naïve Bayes performance is clearly naively orientated towards classifying slots, obtaining very high correction, insertion, deletion and, SER records, clearly representing the worst option. The best results are achieved by J48, obtaining an identical SER than CART, but producing a better f-measure, although none of these methods present acceptable results.

For these experiments, the overall correction amounts are slightly above the baseline, as can be seen in Table 4.2. This zone is mainly composed by elements that resemble fluent words, representing fluency onset, which hardens the classification task, *e.g.*, in cases where the repair does not resemble the reparandum. Consequentially, results for this region are the poorest among the performed binary classifications.

4.3.4 Disfluency Detection

This subsection comprises the outcomes of the binary detection of the disfluent zone as a whole, which are displayed in Tables 4.9 and 4.10.

For the alignment experiments, achieved including filled pauses and fragments as features, CART represents the best option for reliability, obtaining the best precision and SER records, but also lower

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	414	210	1629	66,3	20,3	31,0	90,0	
	J48	476	297	1567	61,6	23,3	33,8	91,2	
	LR	320	109	1723	74,6	15,7	25,9	89,7	0,75
	MP	490	247	1553	66,5	24,0	35,3	88,1	0,76
	NB	604	1487	1439	28,9	29,6	29,2	143,2	0,72
Recognition	CART	260	123	1807	67,9	12,6	21,2	93,4	
	J48	305	189	1762	61,7	14,8	23,8	94,4	
	LR	152	69	1915	68,8	7,4	13,3	96,0	0,69
	MP	322	183	1745	63,8	15,6	25,0	93,3	0,69
	NB	344	805	1723	29,9	16,6	21,4	122,3	0,67

Table 4.10: Detailed performance analysis on predicting the disfluent region as a whole while excluding FP and FRG as features.

recall than the remaining approaches. In terms of recall, Multilayer perceptron stands as the best option, however, the corresponding precision remains well bellow the one obtained by either J48 or Logistic Regression, and moreover the SER metric stresses that Multilayer Perceptron obtains a higher error amount than both these approaches. J48 stands as a less error prone approach than Multilayer Perceptron for obtaining increased recall, obtaining a better f-measure record than the remaining methods. In this case, Logistic Regression represents a half way choice between precision and recall, producing the 2º best records for precision, f-measure and, SER, while producing timely results.

For the alignment data experiments achieved excluding filled pauses (FPs) and fragments (FRGs), both CART and J48 suffer a major loss of precision, showing slightly more pronounced contours in the J48 case. Logistic Regression achieves a combination of: lowest precision, lowest insertion, highest deletion, and the most severe recall degradation. The corresponding SER value remains below the one achieved by the tested decision trees approaches, but still not exceeding half a unit. Multilayer Perceptron performs best, achieving better precision and recall than the decision trees, performing best in almost all performance metrics. Logistic Regression achieves better precision, but suffers substantially the same ratio for recall, resulting in a higher SER record than Multilayer Perceptron. Additionally, Multilayer Perceptron also gets a considerably higher correction amount, representing the most balanced approach among the tested ones.

In what regards the results obtained using recognition data while including FPs and FRGs, the deletion rates suffer sharp increases, followed by less sharper insertion increases. The correction rates are also severely affected with the removal of these features. Multilayer Perceptron stands out as the best option, achieving a slightly lower precision level in comparison to the best records, together with the best results for recall (except Naïve Bayes), f-measure and, SER. For precision, Logistic Regression stands out as the best option, however this approach produces a much inferior recall than CART and J48, resulting in a much lower f-measure record. The decision trees (CART, J48) perform considerably worse than Multilayer Perceptron, obtaining a better f-measure than Logistic Regression, but worse SER

		Cor	Ins	Del	Prec	Rec	F	SER
Alignment	CART	1473	199	2377	88.1	38.3	53.4	66.9
	J48	1499	244	2351	86.0	38.9	53.6	67.4
	LR	1476	263	2374	84.9	38.3	52.8	68.5
	MP	1495	416	2355	78.2	38.8	51.9	72.0
	NB	1841	5427	2009	25.3	47.8	33.1	193.1
Recognition	CART	513	189	3201	73.1	13.8	23.2	91.3
	J48	564	344	3150	62.1	15.2	24.4	94.1
	LR	493	185	3221	72.7	13.3	22.4	91.7
	MP	632	543	3082	53.8	17.0	25.9	97.6
	NB	896	2307	2818	28.0	24.1	25.9	138.0

Table 4.11: Detailed performance analysis on overall binary performances obtained while including FP and FRG as features.

		Cor	Ins	Del	Prec	Rec	F	SER
Alignment	CART	513	189	3201	73.1	13.8	23.2	91.3
	J48	564	344	3150	62.1	15.2	24.4	94.1
	LR	493	185	3221	72.7	13.3	22.4	91.7
	MP	632	543	3082	53.8	17	25.9	97.6
	NB	896	2307	2818	28	24.1	25.9	138
Recognition	CART	192	85	2485	52,5	6.9	12.3	96.1
	J48	186	171	2466	39,5	6.7	11.5	99.5
	LR	140	54	2567	55,3	5.1	9.3	96.9
	MP	215	117	2455	50,2	7.8	13.5	96.5
	NB	399	2102	1973	13,8	14.4	14.1	161.6

Table 4.12: Detailed performance analysis on overall binary performances obtained while excluding FP and FRG as features.

than either Logistic Regression And Multilayer Perceptron.

The results of the recognition data experiments performed excluding FPs and FRGs, reveal lower losses in comparison to the alignment data case. In this case, CART performs better than J48, producing less error, but also considerably lower recall. Logistic regression obtains roughly half the recall felt in the remaining cases, together with the highest precision record, but the corresponding SER metric denounces high error propensity. Naïve Bayes presents an enormous insertion rate, resulting in very high error amounts. Multilayer Perceptron performs best, achieving a higher number of corrections than when FPs and FRGs are included, followed by a considerable increase in recall. In this case the results of the removal of these two features are particularly evident in the outcomes of precision, resulting in linear losses for all methods.

4.3.5 Overall Binary Performance

This subsection analyses method performances relative to all the binary classifications. The data exposed in Table 4.11 is achieved by adding the results of all classifications, while using only: corrections, insertions, and deletions for parameter calculations. The following analysis pretends solely to compare the performance of each method in a general perspective, performing the same calculations as in the binary classifications, while using a summed version of the results of the binary classifications in terms of correct, insertion, and deletion records, each individually.

For the alignment data experiments, performed while including filled pauses and fragments as features, CART and J48 generally achieve the best records. CART accounts for better precision and SER records, while J48 is better in terms of recall and f-measure. Logistic Regression achieves similar results to CART and J48 although slightly lower in therms of precision and SER, presenting a similar recall value as CARTs. Apart from Naïve Bayes, Multilayer Perceptron achieves the lowest precision results, contrasting with a very similar recall record to J48. Naïve Bayes also risks much towards the classification of the less prevalent classification element, resulting in very high error levels.

As presented in Table 4.12, for the alignment data the results achieved excluding filled pauses (FPs) and fragments (FRGs) show significant and transversal performance losses in comparison to the results obtained for alignment. The decision trees and Multilayer Perceptron are most affected in terms of precision, in particular J48, resulting in considerably higher losses in comparison to the remaining methods. Logistic Regression suffers the greatest recall loss, together with a much lower precision reduction in comparison to the decision trees. Except for the decision trees the recall reductions were more severe than the ones seen for precision. Apart from Naïve Bayes, all methods experience large correction amounts, coupled with a quite larger deletion increase. The insertions rate is also inflated, but in this case the increase was only slight in contrast to the SER results, a metric which suffered an accented increase. Multilayer Perceptron seems to perform best, however, the corresponding SER shows this approach is much more error prone than both CART and LR. Although the results achieved by CART outperform those of LR, it is remarkable the resemblance of the performance achieved under these conditions, while LR performs consistently briefer in terms of training phase time consumption.

The recognition data results, achieved while including filled pauses and fragments as features, comprise generalized losses in comparison to the corresponding alignment results, showing considerable performance reductions for both f-measure and SER. Regarding the decision trees, CART accounts for better precision and SER records, while J48 is better in terms of recall and f-measure. However, J48 achieves a considerably worse SER record than CART, and also a somehow daunting precision in comparison to CART and LR. Concerning this metric, LR obtains a close record to CART, while performing much briefly in the training phase. Multilayer Perceptron and Naïve Bayes perform very poorly under these circumstances, resulting in the highest SER records and very low precision.

For the recognition data experiments, the removal of FPs and FRGs generates much less abrupt

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	762	71	1899	81.5	27.6	41.2	75.0	
	J48	749	96	1891	77.4	27.1	40.2	76.4	
	LR	763	118	1883	76.5	27.6	40.6	76.7	0.9
	MP	753	99	1891	77.5	27.3	40.3	76.3	0.9
	NB	980	3983	1317	18.1	35.5	23.9	208.7	0.8
Recognition	CART	274	109	2451	64.6	9.9	17.2	94.0	
	J48	290	144	2411	58.1	10.5	17.8	94.7	
	LR	256	83	2475	68.4	9.3	16.3	93.7	0.8
	MP	311	154	2396	59.4	11.2	18.9	94.3	0.7
	NB	522	1972	1994	19.0	18.9	18.9	152.4	0.7

Table 4.13: Detailed performance analysis for a multi-class prediction performed while including FP and FRG as features.

losses than for alignment, although the results achieved including both these features are already fairly low. Logistic Regression achieves the best precision, however the associated recall record suggests this approach might not be adequate for this conditions, especially when compared to the CART result, that presents a more balanced choice, and also lower SER. Additionally CART covers more correct slot occurrences. J48 performs frankly lower than CART in this case. Multilayer Perceptron achieves the highest amount of corrections, a good SER record (similar to J48) and, a high insertion level. It is noticeable that Multilayer Perceptron registers precision increases with the deletion of FPs and FRGs, together with one of the smallest recall losses. The choice of the best approach is now tied between Logistic Regression and CART, the tier will be based on the SER metric which seems to point CART as the less error prone approach, and Multilayer Perceptron the best for slot coverage. As a final point, note that CART obtains more deletions, while Multilayer Perceptron performs better in this indicator, but more insertions are produced.

4.3.6 Multi-Class Classification

This subsection presents results of a multi-class classification experiment, performed accounting for several distinct disfluency related regions, and also words that rely out of disfluent sequences. The results for every slot where summed and used for the calculation of the overall performances present in Table 4.13.

		Cor	Ins	Del	Prec	Rec	F	SER	ROC Area
Alignment	CART	414	156	2213	15.0	23.9	90.7	0.719	
	J48	430	249	2159	15.6	23.8	93.4	0.683	
	LR	359	143	2285	13.0	21.2	92.2	0.754	0.9
	MP	400	117	2249	14.5	23.6	89.8	0.767	0.9
	NB	741	4312	1469	26.8	17.7	229.2	0.690	0.8
Recognition	CART	192	85	2485	6.9	12.3	96.1	0.606	
	J48	186	171	2466	6.7	11.5	99.5	0.594	
	LR	140	54	2567	5.1	9.3	96.9	0.690	0.8
	MP	215	117	2455	7.8	13.5	96.5	0.689	0.7
	NB	399	2102	1973	14.4	14.1	161.6	0.659	0.7

Table 4.14: Detailed performance analysis for a multi-class prediction performed while excluding FP and FRG as features.

		Cor	Ins	Del	Prec	Rec	F	SER
Alignment	ip	271	82	449	76.8	37.6	50.5	73.8
	interregnum	366	12	1	96.8	99.7	98.3	3.5
	reparandum	19	33	937	36.5	2.0	3.8	101.5
	repair	106	46	614	69.7	14.7	24.3	91.7
	out	*21682*	*1899*	*71*	*87.7*	*95.9*	*92.1*	*12.5*
	overall performance	762	71	1899	81.5	27.6	41.2	75.0
		Cor	Ins	Del	Prec	Rec	F	SER
Recognition	ip	38	53	663	41.8	5.4	9.6	102.1
	interregnum	209	68	38	75.5	84.6	79.8	42.9
	reparandum	6	4	1113	60.0	0.5	1.1	99.8
	repair	21	25	678	45.7	3.0	5.6	100.6
	out	*18152*	*2451*	*109*	*88.1*	*99.4*	*93.4*	*14.0*
	overall performance	274	109	2451	64.6	9.9	17.2	94.0

Table 4.15: Detailed Multi-class classification CART alignment results obtained while including FP and FRG as features.

For the alignment data experiments, performed while including filled pauses and fragments as features, the best results are achieved by CART, obtaining the best precision, recall and, SER records. Logistic Regression shares the best recall record with CART, however, the corresponding performance for precision and SER remains bellow those obtained by using the latter approach. Multilayer Perceptron also performs consistently worse than CART in this case, as does J48.

For alignment, discarding filled pauses (FPs) and fragments (FRGs) results in large losses for all methods. CART suffers a severe precision loss followed by a less sharper recall decay, the corresponding SER also denounces a high performance impact related with the usage of raw recognition transcripts. J48 obtains a worse precision and SER, showing this method performs lower than CART. CART obtains a better precision and SER, but also a lower correction amount. Multilayer Perceptron achieves a worse SER value than CART. Logistic Regression achieves the lowest error, but also the lowest correction amount. The results show that Multilayer Perceptron achieves a smaller number of corrections in comparison to J48, also a proportionally higher amount of deletions. In sum, for these conditions CART represents the best approach for increased precision, and MP is best for increased recall, performing

Classified as	ip	interregnum	reparandum	repair	outside disf
ip	271	0	19	5	425
interregnum	0	366	0	0	1
reparandum	58	0	19	14	865
repair	0	3	3	106	608
outside disfluency	24	9	11	27	21682

Table 4.16: Multi-class classification result matrix of CART results, obtained using forced alignment data and including FP and FRG as features.

better in terms of SER than J48, while at the same time obtaining more corrections.

For the experiments that used recognition data, while including filled pauses and fragments as features, Logistic Regression presents solid results, performing well in the presence of noisy data, obtaining the lowest SER, good precision and also a larger ROC area in comparison to Multilayer Perceptron. In this case CARTs and Multilayer Perceptron share the best option, whereas the former presents better SER and precision while the latter increased correct slot classifications and higher recall values. In comparison to J48, Multilayer Perceptron presents a higher insertion rate, but this is compensated by the corresponding lower deletion values. Naïve Bayes remains the worst option, presenting a much higher SER record than the remaining approaches.

For recognition, the removal of FPs and FRGs results in much less abrupt recall losses than for alignment, although the results fall by half and rely below 10 units. The precision metric suffers dilated losses in a generalized way. J48 obtains the best SER record, but Multilayer Perceptron achieves the best precision, recall and, corrections amount, achieving the best slot oriented performance. Logistic regression presents a balanced choice, but still weaker than either Multilayer Perceptron or CART. CART outperforms J48, obtaining a better correction rate, followed by a well bellow amount of insertions, which does not reflect much in the corresponding deletions, and also a higher amount of corrections. The results corresponding to Naïve Bayes are somehow very poor, in the sense that this approach takes a huge amount of error, resulting in very high rates for corrections and insertions.

Table 4.15 comprises results for the multi-class experiments achieved using CART, for both forced aligned material and raw ASR transcriptions. From all the structural elements related to disfluency, the interregnum region is by far the easiest to detect for both data conditions. This behavior was expected since information about FPs and FRGs is being provided as features, which are known to strongly populate this zone. The reparandum region presents very low classification outcomes, specially when facing recognition data, presenting a very low recall in both cases.

Detecting the interregnum produces the worst classification results, which is expected, due to the resemblance the elements contained present to fluent material. The surrounding regions to an interruption point are often referred in the literature as containing good cues for detecting disfluencies, which is attributed to corresponding characteristic contours known to populate the surrounding word regions.

For the alignment data, the results for this element generate good f-measure and SER records, which is attributed to the fact that the interruption point is often followed by FPs, and sometimes preceded by FRGs, for which information is included as features.

The corresponding results for recognition reveal sharp performance decreases. The repair region presents considerably better results than the reparandum, obtaining more corrections and less deletions, but the performance achieved is still quite weak. For detecting the repair, considerably high precision is achieved, resulting in marked recall losses. In order to improve the recall rate for classifying this zone, a more deep word context analysis is needed. The corresponding results for recognition show marked reductions in both recall and precision, presenting an SER exceeding 100 units.

In Table 4.16 the line concerning the elements outside a disfluency refers to elements that were not considered one of the 4 possible structural elements of a disfluency, and correspond to non-slots. The matrix present in this table concerns the alignment results achieved using CART, showing the interregnum region is the easiest to detect producing only one insertion. Follows the interruption point region, also presenting a high number of corrections, but also a high number of substitutions, pointing out that an improved representation strategy can improve the detection of this zone. The next successful classification zone is the repair, presenting considerably higher perplexity than the one seen for the interruption point. The classification of other words inside disfluency presents the highest perplexity, showing more propensity towards deciding for the interruption point region than for in-disf or repair, but still the 'outside disf' category massively retains classifications, since these words might resemble words that occur outside disfluencies. The majority of classifications are oriented towards elements that are placed "outside of a disfluency", which is the most common situation in the corpus.

The performance for each of the individual structures is even better than performing each one separately, but the result matrix shows that the results are still much influenced by the number of deletions. The overall performance is affected by the low detection performance for the reparandum words, because most of such words are in fact fluent and thus difficult to distinguish from words outside of a disfluency (Nakatani and Hirschberg, 1994; Shriberg, 2001).

4.3.7 Summary of Results

The present subsection comprises a summary of results obtained in the previously described experiments. For the alignment data binary experiments, Logistic Regression outperforms Multilayer Perceptron, performing better while risking roughly the same, showing a tendency towards producing the best results when facing noisy data. These methods presented a clear tendency towards performing better when facing recognition data while either exempting filled pauses and fragments as features or not, which are known to strongly characterize the disfluent region. Logistic Regression tends to achieve

close results to the decision trees, while requiring far less time. However, for the repair region, Logistic Regression does not perform acceptably against the recognition data, Multilayer Perceptron does not run any better in this case. For the detection of the disfluent region as a whole, which is a very general classification in terms of heterogeneity, Logistic Regression reveals comparable performance to the Decision Trees.

Results for the overall binary classifications point the decision trees as the most reliable approach, revealing that CART's provide better precision and SER, while J48 provides better recall and f-measure. Logistic Regression produces similar results to both CART and J48, although slightly lower. Multilayer Perceptron results in much lower precision than both the decision trees and Logistic Regression. Naïve Bayes doesn't seem sensitive to this type of data, since it performs similarly for both alignment and recognition, resulting in very low precision and SER ratings. As for the alignment, Multilayer Perceptron proved to be better for both data-set conditions. For recognition data, Logistic Regression and Multilayer Perceptron tend to provide better performance. In line with J48, Multilayer Perceptron tends to risk much towards the classification of the less prevalent element, resulting in high recall but reduced precision, specially for the alignment data. Multilayer Perceptron tends to perform better when facing noisy data, achieving good results for detecting the whole disfluent region.

The multi-class classifications reveal conflicting results regarding the binary classifications. For alignment, CARTs provide the best results, obtaining the best records for insertions, substitutions, precision, f-measure and, SER. In this case, Logistic Regression performs similarly to Multilayer Perceptron, and better than J48. For recognition, CART and Logistic Regression achieve the best records, revealing once again the latter's potential to perform well towards noisy data, while producing timely results. Naïve Bayes consistently obtains the poorest results showing a consistent tendency towards slot classification, generating very high insertion values in relation to the other approaches. This denounces a hight level of risk acceptance, leading to high recall values, and also very high slot error rates, resulting in drastic performance reductions. The build time consumed by Naïve Bayes tends to decrease proportionally to the increase in the number of slots, as seen in the multi-class classifications results, which also seems to support the view that Multilayer Perceptron is very conductive to risk acceptance. However this is not the case for the interregnum region while using the alignment data, for which Naïve Bayes was the second most costly approach in therms of time consumption. In general the decision trees suffer accentuated performance losses when facing recognition data. This relates to a corresponding disjunctive nature, that leads to less tolerance to strange elements that income from the ASR, which may resemble learned elements. The repair results confirm this fact, since for recognition the best methods for dealing with reliable information are the decision trees. Note the repair region is mostly composed of fluent elements.

In general CARTs and J48 achieve the best records when facing forced aligned material, generally achieving better precision records, while Logistic Regression and Multilayer perception tend to perform

	Feature	inDisf	ip	int.	repair	All
1	$num.syl.w_0$	•••		•••••		•••••
2	$duration.racio.w_0.w_1$	•••••	•••••	•••••	•••••	•••••
3	$equality.w_0.w_{+1}$	•••••	•••••		•••••	••••
4	$b.sil.racio.w_0.w_{+1}$		•••••	•••••	•••••	•••••
5	$p.med.ratio.w_0.w_{+1}$	••••	••	•••••	••••	•••••
6	$conf.w_0$	•••••	••••	••••	••••	•••
7	$equality.w_0.w_{-1}$	•••••	••••	•••	•••••	••••
8	$e.med.ratio.w_0.w_{+1}$	••••	•••	••••	••	••••
9	$num.phones.w_0$	••••	••	••••		••••
10	$e.slope.w_0w_{+1}(RR)$	•••	•••••	•	••	••
11	$b.sil.comp.w_0.w_{+1}(>)$	•••	••••	••••	•••	•••
12	$conf.w_{+1}$		••	•	••••	••••
13	$b.sil.comp.w_0.w_{+1}(<)$	••••	•••	••••	••	•••
14	$e.slope.w_0w_{+1}(FF)$	•••	••••	••	••••	•
15	$p.slope.w_0w_{+1}(R-)$	•••	•	•••	••••	•••
16	$b.sil.comp.w_0.w_{+1}(=)$	••••	••••	•••	••	••
17	$p.slope.w_0w_{+1}(RF)$	••	••	•••	•	•••
18	$pslopes : RF_{cw,fw}$	•	•	••	•••	••
19	$e.slope.w_0w_{+1}(RF)$	•	•••	•	•	•
20	$e.slope.w_0w_{+1}(R-)$				•••	

Table 4.17: Top 20 most influent features for forced alignment obtained while discarding FP and FRG.

better when facing noisy data, and also when facing a poorer set of features. Although based on these metrics the performance of the multi-class experiments seems to be lower than the ones achieved for the disfluent regions individually, an analysis performed in the following section on the same results reveals that the performance achieved for each individual region is superior to accounting for the region individually in binary classifications. Regarding the binary classifications, we observe a clear trend in the Decision Tree's behavior, CARTs produce better precision and SER records, while J48 provides better recall and higher correction rates. This trend tends to prevail while facing both forced alignment, and automatically recognized data types. For alignment, CARTs tend to only classify slots in the presence high assurance levels. In contrast, J48 risks much more towards data of this kind. For the repair region, J48 seems more accurate than CARTs when facing data with few cues such as the repair region, on which J48 performs slightly better for both recognition, and alignment data.

4.3.8 Feature Impact Analysis

In order to access the influence of the adopted features on the classification, the approach relies on the analysis of the tree generated by CARTs. In general, the set of features that proved most informative for cross-region identification encompasses word duration ratios, word confidence score, silent ratios, and pitch and energy slopes. Moniz et al. (2009, 2011a) 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

	Feature	inDisf	ip	int.	repair	All
1	$fragment$	•••••	•••••	•••••	•••••	•••••
2	$filled.p$	•••••	•••••	••••	•••••	•••••
3	$num.syl.w_0$			•••••		•••••
4	$dur.ratio.w_0.w_{+1}$	••••	•••••	•••	••••	••••
5	$equality.w_0.w_{+1}$	•••••	•••••	•	•••••	••••
6	$conf.w_0$	•••••	••••	•••••	••••	•••••
7	$b.sil.ratio.w_0.w_{+1}$		•••	•••••	•••••	••••
8	$p.med.ratio.w_0.w_{+1}$		•••	••••	••••	••••
9	$conf.w_{+1}$	•••	••••	••••	••••	••••
10	$e.med.ratio.w_0.w_{+1}$	••••	••••	•••	•	•••
11	$filled.p.w_{+1}$	•••	••••		••	•••
12	$equality.w_0.w_{-1}$	••••	••••		••••	••
13	$p.slope.w_0.w_{+1}(F-)$			••••		•••
14	$num.phones.w_0$	••••	••	••••		••
15	$e.slope.w_0.w_{+1}(RR)$	••••	•••		••	
16	$p.slope.w_0.w_{+1}(FR)$		•	•••	•••	•••
17	$p.slope.w_0.w_{+1}(R-)$	••	••	•••	•••	•••
18	$e.slope.w_0.w_{+1}(FF)$	••	••		•••	•
19	$p.slope.w_0.w_{+1}(FR)$	•	•••			
20	$b.sil.comp.w_0.w_{+1}(=<)$	•••	•	••	•	

Table 4.18: Top 20 most influent features for forced alignment, obtained while considering FP and FRG.

pauses.

As shown in Table 4.17, for the experiments that excluded filled pauses (FPs) and fragments (FRGs), features such as $num.syl.w_0$, $num.phones.w_0$, $p.med.ratio.w_0.w_{+1}$ *and,* $dur.comp.w_0.w_{+1}$, proved to be more useful for the identification of the interregnum, whereas energy slopes were most suited for identifying the interruption point, although these tend to be confounded with the repair. For detecting the repair while excluding these features, the most reliable cues are, $conf.w_0$, $e.slope.w_0.w_{+1}$, $p.slope.w_0.w_{+1}$ and, $equality.w_0.w_{-1}$. The best features for discriminating the disfluent region are, $conf.w_0$, $dur.ratio.w_0.w_{+1}$ and, word equality comparisons. For detecting all disfluency related regions simultaneously (reparandum, interruption point, interregnum, repair), including a slot for words outside of disfluency and excluding FPs and FRGs, the most relevant features are $num.syl.w_0$, $dur.ratio.w_0.w_{+1}$, $b.sil.ratio.w_0.w_{+1}$ and $p.med.ratio.w_0.w_{+1}$.

The best features for the experiments performed including FPs and FRGs are displayed in Table 4.18. For the interruption point task, performed including FPs and FRGs as features, the best features in terms of zone differentiation potential are, $p.slope.w_0.w_{+1}$ and $e.slope.w_0.w_{+1}$, although the latter also strongly characterizes the repair region. Another feature that is also very representative of both these regions is $equality.w_0.w_{+1}$. In terms of distinguishable characteristics between zones, the best features for detecting the interregnum when including FPs and FRGs are: $num.syl.w_0$, $num.phones.w_0$ and, $p.slope.w_0.w_{+1}$. The best features for detecting the repair region, when including FPs and FRGs are, $e.slope.w_0.w_{+1}$, and $b.sil.ratio.w_0.w_{+1}$. For detecting the disfluent region in it's totality, while in-

	time(sec.)		overall perf.		detailed slot performance							
	train	test	acc.	kappa	cor	ins	del	prec	rec	F	SER	ROC
ZeroR	0.1	2.1	98.42	0.00	0	0	388					0.50
Simple CART	1257	1.7	98.82	0.55	179	80	209	69.1	46.1	55.3	74.5	
J48	1800	1.9	98.87	0.60	217	107	171	67.0	**55.9**	61.0	71.6	
Logistic Regression	33	2.1	98.74	0.47	139	59	249	**70.2**	35.8	47.4	79.4	0.98
Multilayer Perceptron	3516	7.9	98.71	0.55	201	129	187	60.9	51.8	56.0	81.4	0.97

Table 4.19: Performance Analysis on Predicting *filled pauses*

cluding FPs and FRGs, the most relevant features are, $conf.w_0$ and $equality.w_0.w_{+1}$. Features such as $dur.ratio.w_0.w_{+1}$, $e.med.ratio.w_0.w_{+1}$ and, $equality.w_0.w_{-1}$, are also highly relevant. In terms of weight the best features for the multi-class experiment when including FPs and FRGs are $num.syl.w_0$ and $conf.w_0$, but $dur.ratio.w_0.w_{+1}$, $b.sil.ratio.w_0.w_{+1}$ and, $p.med.ratio.w_0.w_{+1}$, are also highly relevant.

4.4 Filled Pause Detection

This subsection studies the viability of automatically identifying filled pause events (FP), based on both the existing audio segmentation given by the recognizer, and additional prosodic features. The study of filled pauses (FPs) is motivated by the fact that these are very frequent in several European Portuguese speech domains analyzed in previous work, and also because these are generally described in the literature as the most frequent event.

The lexicon of the ASR AUDIMUS (Meinedo et al., 2003; Meinedo, 2008) was recently upgraded with entries containing possible phonetic sequences for a FP. Such additional entries made it possible to automatically detect these structures. Experiments used the same conditions used in previous sections, except for the exclusion of FPs and fragments from the feature set, and the abandonment of Naïve Bayes, since this approach consistently performed poorly in the previous experiments. The tests in this section exclude fragments (FRGs) as cues, in order to explore the impact of the proposed prosodic features on the FP detection task.

Table 4.19 summarizes the classification results, which are analyzed using metrics described in Section 4.1. The second and third columns of this table report on the time (seconds) taken for training and testing the models, revealing that Logistic Regression is considerably faster on the training phase in comparison to the other methods, performing 38 times faster than any other approach. The values presented in the remaining columns consider only slots, which in this case are FPs, corresponding to more meaningful performance metrics, as described in Section 4.1. The high accuracy values present in Table 4.19 point that the data is highly unbalanced. In fact regular words correspond to 98.42% of the total

	Cor	Ins	Del	Prec	Rec	F	SER
Current ASR system	223	146	140	60.4	**61.4**	60.9	78.8
J48	217	107	171	**67.0**	55.9	61.0	**71.6**

Table 4.20: Current ASR Results.

events, which poses increased difficulty for the classification task. J48 is clearly the best suited method for this task, achieving simultaneously the highest percentage of overall correct classifications and the best performance when considering slots only. This method generates a fairly low amount of insertions and deletions, while achieving a significantly higher number of correct instances than the remaining methods. Logistic Regression obtains a contrasting highest precision, and lowest recall, resulting in a very low f-measure record. The tree based approaches (CART / J48) consume approximately half the time of Multilayer Perceptron (MP) for model training, while achieving better performance.

4.4.1 Feature Impact Analysis

In order to access the influence of the adopted features on the classification of filled pauses (FP), the tree generated by J48 is analyzed. A total of 498 leaves are produced, but the top most decisions in the tree lead to the set of most informative descriptors. Following are the features sorted by order or relevance for the classification of FPs: i) confidence score of the current word; ii) current word is composed of a single phone and is lengthier than the following word and; iii) current word has adjacent silent pauses, plateau pitch contours; and iv) current word maximum energy. The obtained findings are inline with work for English, reported by O'Shaughnessy (1992) and Shriberg et al. (1997), a.o., since adjacent silent pauses, plateau pitch contours, and constant energy values stand out as the most discriminant features. The fact that European Portuguese shares with English those properties represents a contribution more to cross-language understanding, than for the goals pursued in the present work. What this study adds, is the crucial importance of two features: the confidence level and the number of phones.

4.4.2 ASR Approach Comparison

This section comprises results and analysis for comparing the best outcomes of the proposed approach for detecting filled pauses (FP), achieved with J48, and the currently implemented ASR approach, aiming at assessing whether an approach that uses prosodic features may be useful for extending our current system. The results achieved represent a parallel way to assess the prediction of FP in the ASR system. The lexicon of the ASR AUDIMUS (Meinedo et al., 2003; Meinedo, 2008) was recently upgraded with entries containing possible phonetic sequences for a FP. Such additional entries made it possible

to automatically detect these structures. Experiments use the same data subset used in the previous sections, and Table 4.20 comprises the resulting data.

Results are quite similar in terms of f-measure. The precision and recall metrics show a much more dilated discrepancy, still the current ASR system tends to perform similarly in terms of both precision and recall. J48 achieves a significantly higher precision, while the current ASR approach better recall, revealing a higher propensity for error in the latter case, which becomes obvious based on the analysis of the corresponding SER. It is also noteworthy the impact of including FPs in the lexicon with alternative pronunciations achieved by the ASR. Results suggest that combining both approaches may lead to better performances.

4.4.3 Filled Pause Conclusion

This section presented a number of experiments aimed at automatically detecting filled pauses (FP) in a corpus of university lectures, using four different machine learning methods: CART, J48, Logistic Regression, Multilayer Perceptron. The aim relies on assessing how well a system relying on prosodic features can complement or outperform the current ASR FP detection system, which is based on adding possible phonetic sequences of FPs to the lexicon of the recognizer. The Experiments described in 4.3 assumed that information about FPs was previously given by a manual annotation. The experiments presented in this section represent a step forward automatically detecting disfluencies, since the performance for automatically calculating FPs information is now given. Although both approaches perform quite similarly in terms of f-measure, the SER is almost 7% (absolute) better for J48. The best results are achieved by J48, inline several literature forecasts (Shriberg, 1994; Nakatani and Hirschberg, 1994; Shriberg et al., 1997).

Several variables difficult the process of comparing results with further work, namely: corpora differences, languages, domains, and evaluation setups. From a linguistic point of view, Portuguese FPs are often ambiguous with very frequent functional words. The filled pause “aam” is also ambiguous with verbal forms due mostly to possible vowel reduction or deletion in the word final position, as well as with acronyms. As another example note the filled pause “mm” may be recognized as the article “um” / ’a’, or the cardinal number “um” / “one”.

From a state of the art perspective, this work does not include phoneme related information, part-of-speech, syntactic and other multimodal information. This study shows that prosodic features alone produce results comparable to accounting for these phenomena using both language (syntactic) and acoustic models.

5 Conclusion

The work described in this dissertation presents a number of experiments focusing: on the automatic identification of disfluent sequences, on distinguishing between their structural elements, and on the filled pause (FP) detection task. Different machine learning methods have been tested, using a corpus of university lectures in European Portuguese.

Initial experiments target the detection of all zones related to disfluency, performing comparisons between alignment and recognition data results, and also for the impact of FPs and fragments (FRG) under both data conditions. To the best of our knowledge, this is the first work that automatically identifies disfluencies and their structural elements for a Portuguese corpus using a machine learning approach and using mostly prosodic cues, and represents an important step in the development of this kind of systems for our language. The results for these experiments using alignment data and including filled pauses and fragments as features, show that the performance achieved for detecting words inside of disfluent sequences when FP and FRG are used as a features, is about 91% precision and 37% recall, corresponding to the CART results. Multilayer Perceptron presents the best approaches for increased recall, while Logistic Regression achieves the most balanced results, while performing much faster in the building phase. Results for this region achieved by CART, while removing FPs and FRGs from the alignment experiments, show precision lowers to 66% and recall to 20%, respectively, stressing the importances of these cues.

The results of the alignment experiments regarding the detection of disfluent sequences, while excluding FPs and FRGs, suggest that CARTs and Logistic Regression can achieve similar results. While CART tends to achieve better precision, Logistic Regression accounts for increased recall. Logistic Regression presents the best choice in terms of computational effort, performing much faster than the remaining classification approaches on the binary classifications. The best approach for detecting disfluencies while using recognition data and including filled pauses and fragments belongs to Multilayer Perceptron for recall, and Logistic Regression for increased precision. Results for detecting the disfluent region, achieved without filled pauses and fragments and using the alignment data, reveal that Multilayer Perception represents the best approach for both recall and SER. LR achieves good precision but also a very low recall. CARTs seem to outperform Logistic Regression on the FP and FRG disproved exper-

iments, however, the SER obtained by Logistic Regression remains below the records achieved by the decision trees.

Results for the overall binary classifications concerning the alignment data, point the decision trees as the most reliable approaches, revealing that CART's provide better precision and SER, while J48 provides better recall and f-measure. Logistic Regression produces similar results to both CART and J48, although slightly lower. Multilayer Perceptron results in much lower precision than both the decision trees and Logistic Regression. Naïve Bayes doesn't seem sensitive to this type of data, since it performs similarly for both alignment and recognition, resulting in very low precision and SER ratings. As for the alignment, Multilayer Perceptron proved to be better for both data-set conditions.

The multi-class classifications reveal conflicting results regarding the binary classifications. For the alignment experiments that excluded filled pauses and fragments, CARTs provide the best results, obtaining the best records for insertions, substitutions, precision, and also f-measure and SER. For recognition, including filled pauses and fragments, shows that in this case the best choice is J48 for increased correction rates, and Multilayer Perceptron if a higher recall is preferred. Interestingly, better results are achieved for the classification of the distinct disfluency zones than for the binary classifications of these zone individually, showing that accounting for all disfluent regions simultaneously results in enhanced disambiguations.

Concerning the detection of the disfluent parts of a disfluency, the interregnum region is the easiest to detect, maintaining this position even when unreliable transcripts and the FPs and FRGs, resulting in the best classification performance under all condition combinations. Follows the interruption point region, presenting a high number of corrections, but also a high number of substitutions, pointing that an improved representation strategy can improve the detection of this zone. The next successful classification zone is the repair, presenting considerably higher perplexity than one registered for the interruption point. The classification of other words inside disfluency presents the highest perplexity, showing more propensity towards deciding for the interruption point region than for both in-disf or repair, but still the 'outside disf' category massively retains classifications, since these words might resemble words that occur outside disfluencies. The vast majority of classifications are oriented towards elements that relay "outside of a disfluency", the most common situation in the corpus.

The set of features that proved most informative for cross-region identification encompasses word duration ratios, word confidence score, silent ratios, and pitch and energy slopes. The interruption point is best distinguished by energy slopes, whereas features such as the number of phones and syllables per word proved to be more useful for the identification of the interregnum. The proposed feature set generates acceptable results for discriminating the interruption point and repair, resulting in reasonable precision (53% and 64%, respectively), but the recall metric is transversely negatively impacted. In general, the results show that the most expressive features are confidence scores, word duration ratio,

and knowing when words are equal. On the other hand, features like number of syllables, and number of phones have more impact in specific tasks. Features such as pitch shapes, energy slopes and, silence and duration comparisons, proved to be very informative.

As for the experiments aimed at detecting FPs, the best results are achieved by J48, inline with several literature previews (Levelt, 1989; Liu and Shriberg, 2007; Liu et al., 2006; Makhoul et al., 1999). In these experiments Naïve Bayes is abandoned, since it consistently performed poorly on previous classifications. In order to access how well a system relying on prosodic features could complement or outperform the current ASR filled pause detection system, the best approach (J48) is compared to the in-house implemented solution, revealing that J48 accounts for better precision, while the ASR solution produces better recall. The f-measure values achieved are quite similar, but the SER metric shows the proposed approach is almost 7% absolute better, which might be a more appropriate metric. The main outcome of the FP experiments concerns the fact that prosodic features by themselves do have a strong impact in this task, comparable to accounting for these phenomena in both language and acoustic models. Following are the features sorted by order or relevance for the classification of FPs: i) confidence score of the current word; ii) current word is composed of a single phone and is lengthier than the following word; and iii) current word has adjacent silent pauses, plateau pitch contours; and iv) current word energy maximum. This comes inline with findings for English, reported by O'Shaughnessy (1992); Shriberg et al. (1997). From a state-of-the-art perspective, this study does not include phone related information, part-of-speech, syntactic and other multimodal information. The present study demonstrated that prosodic features by themselves do have a strong impact in the task of detecting FPs, producing comparable results to accounting for these phenomena in both language and acoustic models.

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. Additionally, we are planning a similar work for distinguishing between disfluency locations and punctuation marks. In what concerns FPs, in the future the intent relies in combining this proposal with the current ASR system for better identifying FPs in European Portuguese, and also in the exploration of additional lexical features, as an attempt to improve the FP detection task. Future experiments will apply the proposed system for tasks such as charactering speaking styles and even the speaker. There are several machine learning approaches that have performed well on previous work, such as Conditional Random Field (CRF). Therefore, there is also the intent of experimenting other classification approaches such as Conditional Random Fields.

Bibliography

Adell, J., Bonafonte, A., and Mancebo, D. E. (2008). On the generation of synthetic disfluent speech: local prosodic modifications caused by the insertion of editing terms. In *INTERSPEECH*, pages 2278–2281. ISCA.

Allwood, J., Nivre, J., and Ahlsén, E. (1990). Speech management - on the nonwritten life of speech, nordic. *Journal of Linguistics*, pages 3–48.

Batista, F., Moniz, H., Trancoso, I., Mamede, N., and Mata, A. (2012). Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. *Journal of Speech Sciences*, 2(2):115–138.

Bear, J., Dowding, J., and Shriberg, E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, ACL '92, pages 56–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

Benus, S., Enos, F., Hirschberg, J., and Shriberg, E. (2006). Pauses in deceptive speech. *Aphasiology*, 15(6):571–583.

Candea, M. (2000). *Contribution à l'Etude des Pauses Silencieuses et des Phénomènes dits « d'Hésitation » en Français Oral Spontané – Etude sur un corpus de récit en classe de Français*. PhD thesis, Université de Paris III – Sorbonne Nouvelle.

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

Clark, H. H. and Tree, J. E. F. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111.

Clark, H. H. and Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, 37(3):201–242.

Consortium, L. D. (2004). Simple metadata annotation specification (mde).

Dannélls, D. (2007). Disfluency detection in a dialogue system. Technical report, GSLT, Sweden.

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

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

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, 47:27–52+.

Gormana, K., Lai, C., Yuan, J., and Liberman, M. (2000). Acoustic correlates of cross-linguistic disfluency perception.

Goto, M., Itou, K., and Hayamizu, S. (1999). A real-time filled pause detection system for spontaneous speech recognition. In *Proceedings of Eurospeech '99*, pages 227–230. ISCA.

Gravano, A., Levitan, R., Willson, L., Benus, S., Hirschberg, J., and Nenkova, A. (2011). Acoustic and prosodic correlates of social behavior. In *Interspeech 2011*, Florence, Italy.

He, Y. and Young, S. (2004). *Robustness Issues in a Data-Driven Spoken Language Understanding System*, chapter Robustness Issues in a Data-Driven Spoken Language Understanding System. HLT/-NAACL 2004 Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing, Boston, MA, USA.

Heeman, P. and et al. (1994). Detecting and correcting speech repairs. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Heeman, P. A. and Allen, J. F. (1999). Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25:527–571.

Heike, A. (1981). A content-processing view of hesitation phenomena. Language and Speech, no. 24.

Hieke, A. E. (1981). A content-processing view of hesitation phenomena. *in Audiology & Speech-Language Pathology*, (no. 24).

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

Honal, M. and Schultz, T. (2005). Automatic disfluency removal on recognized spontaneous speech - rapid adaptation to speaker-dependent disfluencies. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.

Jones, D., Wolf, F., Gibson, E., Williams, E., Fedorenko, E., Reynolds, D., and Zissman, M. (2003). Measuring the readability of automatic speech-to-text transcripts. In *Proc. of Eurospeech*, pages 1585–1588.

Kahn, J. G., Ostendorf, M., and Chelba, C. (2004). Parsing conversational speech using enhanced segmentation. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 125–128, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kim, J. and Woodland, P. C. (2001). The use of prosody in a combined system for punctuation generation and speech recognition. In *Proc. of Eurospeech*, pages 2757–2760.

Lai, C., Gorman, K., Yuan, J., and Liberman, M. (2007). Perception of disfluency: language differences and listener bias. In *Proc. Interspeech-2007*, pages 2345–2348. ISCA.

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

Levelt, W. (1989). *Speaking*. MIT Press, Cambridge, Massachusetts.

Liu, Y. (2003). 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. Association for Computational Linguistics.

Liu, Y. and Shriberg, E. (2007). Comparing evaluation metrics for sentence boundary detection. In *Proc. of the IEEE international conference on Acoustics, speech and signal processing*, pages IV–185 – IV–188, Honolulu, Hawaii. IEEE.

Liu, Y., Shriberg, E., Stolcke, A., Hillard, D., Ostendorf, M., and Harper, M. (2006). Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1526–1540.

Makhoul, J., Kubala, F., Schwartz, R., and Weischedel, R. (1999). Performance metrics for information extraction. In *Proc. of the DARPA Broadcast News Workshop*, Herndon, VA.

Mata, A. I. (1995). Apresentação preliminar do cpe-faces: um 'corpus de português europeu falado por adolescentes em contexto escolar', para o estudo da prosódia dos estilos de fala.

Mata, A. I. (2000). *Para o Estudo da Entoação em Fala Espontânea e Preparada no Português Europeu*. PhD thesis, F.L. Universidade de Lisboa.

Meinedo, H. (2008). *Audio pre-processing and speech recognition for broadcast news*. PhD thesis, IST, Universidade Tecnica de Lisboa, Lisboa.

Meinedo, H., Caseiro, D., Neto, J. a., and Trancoso, I. (2003). Audimus.media: A broadcast news speech recognition system for the european portuguese language. In Mamede, N. J., Baptista, J., Trancoso, I., and das Graças Volpe Nunes, M., editors, *Proceedings of the 6th International Conference on Computational Processing of the Portuguese Language*, volume 2721 of *PROPOR'03*, pages 9–17, Berlin, Heidelberg. Springer-Verlag.

Moniz, H., Batista, F., Mata, A. I., and Trancoso, I. (2012a). Analysis of disfluencies in a corpus of university lectures. In *ExLing 2012*.

Moniz, H., Batista, F., Trancoso, I., and da Silva, A. I. M. (2012b). Prosodic context-based analysis of disfluencies. In *Interspeech 2012*, volume Vols 1-3, Portland, Oregon. ISCA, ISCA.

Moniz, H., Batista, F., Trancoso, I., and Mata, A. I. (2011a). *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.

Moniz, H., Mata, A., and Trancoso, I. (2008). How can you use disfluencies and still sound as a good speaker? In *Interspeech*, page 1687.

Moniz, H., Mata, A. I., and Trancoso, I. (2011b). A classificação das disfluências como mecanismos de (dis)fluência e os seus contextos prosódicos. In *Textos Seleccionados do XXVI Encontro Nacional da APL. Porto, APL*.

Moniz, H., Mata, A. I., and Viana, C. (2007). On filled-pauses and prolongations in european portuguese. In *INTERSPEECH*, pages 2645–2648. ISCA.

Moniz, H., Trancoso, I., and Mata, A. I. (2009). Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts. In *Interspeech 2009*, Brighton, England.

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

O. Hasegawa S. Hayamizu K. Tanaka K. Itou, T. A. (1999). A japanese spontaneous speech corpus collected using automatic inference wizard of oz system.

O'Connell, D., S. K. (2005). Uh and um revisited: Are they interjections for signaling delay? uh and um revisited: Are they interjections for signaling delay? uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*.

O'Shaughnessy, D. (1992). 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. IEEE Computer Society.

O'Shaughnessy, D. (1994). Correcting complex false starts in spontaneous speech. In *Acoustics, Speech, and Signal Processing*, pages vol.1, I:349 – I:352.

Parlikar, A., Black, A. W., and Vogel, S. (2010). Improving speech synthesis of machine translation output. In Kobayashi, T., Hirose, K., and Nakamura, S., editors, *INTERSPEECH*, pages 194–197. ISCA.

Rodriguez, L. J. and Torres, M. I. (2006). Spontaneous speech events in two speech databases of human-computer and human-human dialogs in spanish. *Language and Speech*, Vol. 49(Issue 3):p. 333.

Savova, G. and Bachenko, J. (2003). *Prosodic features of four kinds of disfluencys*, pages 91–94. In Disfluency in Spontaneous Speech.

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

Shriberg, E. (2001). To "errrr" is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

Shriberg, E., Bates, R., and Stolcke, A. (1997). A prosody-only decision-tree model for disfluency detection. In Kokkinakis, G., Fakotakis, N., and Dermatas, E., editors, *Proc. Eurospeech*, pages 2383–2386. ISCA.

Shriberg, E. and Stolcke, A. (1996). Word predictability after hesitations a corpus-based study. In *ICSLP*, volume vol.3, pages 1868 –1871. ISCA.

Shriberg, E. and Stolcke, A. (2002). Prosody modeling for automatic speech recognition and understanding. In *in Proc. Workshop on Mathematical Foundations of Natural Language Modeling*, volume 138, pages 105–114. Springer New York.

Stolcke, A. and Shriberg, E. (1996). Statistical language modeling for speech disfluencies. In *in Proc. ICASSP*, pages 405–408.

Stouten, F. and Martens, J.-P. (2004). Benefits of disfluency detection in spontaneous speech recognition. In *Cost 278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction, Proceedings*, page 4 pages, Norwich, UK. Department of Electronics and information systems, International Speech Communication Association (ISCA).

Swerts, M., Wichmann, A., and Beun, R.-J. (1996). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30(4):485–496.

Trancoso, I., Martins, R., Moniz, H., Mata, A. I., and Viana, C. (2008). The lectra corpus - classroom lecture transcriptions in european portuguese. In *LREC*, Marrakech, Morocco. European Language Resources Association.

Tsiaras, V., Panagiotakis, C., and Stylianou, Y. (2000). Video and audio based detection of filled hesitation pauses in classroom lectures.

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

Yang Liu, Elizabeth Shriberg, A. S. (2003). Automatic disfluency identification in conversational speech using multiple knowledge sources. *Transactions on Audio, Speech and Language Processing*, 17(7):1263–1278.

Yeh, J.-F. and Yen, M.-C. (2012). Speech recognition with word fragment detection using prosody features for spontaneous speech. *Applied Mathematics and Information Sciences*, vol 6 no. 55:669S–675S.