Towards automatic language processing and intonational labeling in European Portuguese

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

[1]Faculdade de Letras da Universidade de Lisboa,

[2]Instituto de Engenharia de Sistemas e Computadores – Investigação e Desenvolvimento em Lisboa

[3]ISCTE - Instituto Universitário de Lisboa

[4]Instituto Superior Técnico

Abstract

This work describes a framework that encompasses multi-layered linguistic information, focusing on prosodic features (pitch, energy, and tempo patterns), uses such features to distinguish between sentence-form types and disfluency/fluency repairs, and contributes to the characterization of intonational patterns of spontaneous and prepared speech in European Portuguese. Different machine learning methods have been applied for discriminating between structural metadata events, both in university lectures and in map-task dialogues, containing large amounts of spontaneous speech. Results show that prosodic features, and particularly a set of very informative features, are crucial to distinguish between sentence-form types and disfluency/fluency repair events. This is the first work for European Portuguese on both fully automatic processing of multi-layered linguistically description of spoken corpora and intonational labeling.

Prosody, Speech Processing, European Portuguese, Structural Metadata.

1 Introduction

Studies on the prosodic behavior of structural metadata events require prosodic annotations of such events, which are commonly manually performed for small subsets, but almost impractical for extensive data sets. Therefore, an automatic prosodic labeling, flexible enough to account for disfluent structures but also for other linguistic events, becomes of fundamental importance. Moreover, the analysis of prosodic features is also crucial to model metadata events, in particular, and to improve language-processing technologies, in general. The literature has documented a set of phenomena that signal prosodic boundaries in speech (e.g., Frota, 2009; Gussenhoven, 2004; Ladd, 2008; Ostendorf et al., 2008; Rosenberg, 2009; Shriberg, Favre, Fung, Hakkani-Tür, & Cuendet, 2009; Viana, 1987), even though languages may show variations in the use of this set of features (e.g., Hirst & Di Cristo, 1998; Jun, 2005 and references therein). Based on cross-language comparisons and in the more productive phenomena used to delimit boundaries, there are language-independent proposals, such as the one of Vaissière (1983), stating that the set of prosodic features (pause, final lengthening, pitch movements, and energy resets) are amongst the most salient cues. Building on that assumption, algorithms using such features do seem to be transversal in different language processing technologies.

Enriching automatic speech transcripts with structural metadata (Liu, et al. 2006; Ostendorf et al., 2008), namely punctuation marks and disfluencies, may highly contribute to the legibility of the string of words produced by a recognizer. This may be important for so many applications that the speech recognition system pipeline often integrates several other modules, such as audio segmentation, capitalization, punctuation, and identification of disfluent regions. The task of enriching speech transcripts can be seen as a way to structure the string of words into several linguistic units, thus providing multi-layered structured information, which encompasses different modules of the grammar. Different sources of information may be useful for this task, going much beyond the lexical cues derived from the speech transcripts, or the acoustic cues provided by the audio segmentation module (e.g., speech/non-speech detection, background conditions classification, speaker diarization). In fact, one of the most important roles in the identification and evaluation of structured metadata is played by prosodic cues.

Our goal is to study the impact of prosodic information in revealing structural metadata, simultaneously addressing the tasks of recovering punctuation marks and of identifying disfluencies. The former is associated with the segmentation of strings of words into speech acts, and the latter, besides other aspects, also allows the discrimination of potential ambiguous places for a punctuation mark. Punctuating spontaneous speech is itself a quite complex task, further intensified by the difficulty in segmenting disfluent sequences, and in differentiating between those structural metadata events. Annotators of the corpora used in this study report that those tasks are the hardest to accomplish, this problem is evident in the evaluation of manual transcripts, since the attribution of erroneous punctuation marks to delimit disfluent sequences corresponds to the majority of the errors. Furthermore, prosodic cues either for the attribution of a punctuation mark or for the signalling of a repair may be ambiguous, as shown in Batista, Moniz, Trancoso, and Mamede (2012a) and Moniz, Batista, Trancoso, and Mata (2012).

This document is organized as follows: Section 2 reports the related work. Section 3 briefly describes the data and its processing stages. Section 4 describes the process of adding the prosodic layer. Section 5 reports the most recent results on the classification of metadata events. Finally, Section 6 presents some conclusions and future work.

2 Related work

According to Shattuck-Hufnagel & Turk (1996), prosody has two components: firstly, the acoustic correlates and, secondly, their relation to the organizational structure of a language. Detailed analyses have been conducted to describe the properties of the prosodic constituents and their functions (e.g., Beckman & Pierrehumbert, 1986; Bolinger, 1989; Bruce, 1977; Gussenhoven, 2004; Ladd, 2008; Liberman, 1975; Nespor & Vogel, 2007; Pierrehumbert, 1980; Pierrehumbert & Hirschberg, 1990). Since the focus of this work is on the acoustic correlates of structural metadata events, we will briefly comment the higher-level structures described in the literature.

In the Intonational Phonology framework, based on the study of Pierrehumbert (1980) and much subsequent work (e.g., Beckman, Hirschberg, & Shattuck-Hufnagel, 2005; Beckman & Pierrehumbert, 1986; Pierrehumbert & Hirschberg, 1990), the prosodic structure involves a hierarchy of prosodic constituents, encompassing from mora or syllable, the smallest constituents, to intonation phrase or utterance, the largest ones. This hierarchical structure varies in what regards intermediate levels, namely the intermediate intonational phrase (e.g., Beckman & Pierrehumbert, 1986; Frota, 2012; Gussenhoven, 2004; Ladd, 2008; Nespor & Vogel, 2007).

Cross-language studies have also investigated the acoustic correlates that better characterize sentence-like unit boundaries (Vaissière, 1983). Features that are known to characterize higher-level structures, such as pause at the boundary, pitch declination over sentences, post-boundary pitch and energy resets, pre-boundary lengthening, and voice quality changes, are amongst the most salient cues to detect sentence-like units. This set of prosodic properties has been used in the literature to successfully detect punctuation marks and disfluencies. By studying the acoustic correlates of sentence-like units and disfluencies in European Portuguese (EP) we expect to detect higher-level structures of speech, as intonational phrases and utterances.

To inform the automatic process of capturing the prosodic behaviour of sentence-form types in EP, we will briefly review the literature. Declaratives are the most studied sentence type (e.g., Cruz-Ferreira,1998; Frota,2000; Viana, 1987; Viana et al., 2007; Vigário, 1995). The intonational contour generally associated with a declarative is a falling one, expressed as a prenuclear H* (in the first accented syllable), a nuclear bitonal event H+L*, and a boundary tone L%, according to the Tones and Break Indices (ToBI) labelling system (Silverman et al., 1992) adapted for Portuguese. A similar intonational contour is found in wh- questions (Cruz-Ferreira, 1998). By contrast, the contour associated with a yes/no question is a rising one, expressed either as H* H+L* H% or (H) H+L* LH% (the latter proposed by Frota, 2002). Mata (1990) also observes falling contours in yes/no questions in spontaneous speech. As for alternative questions, only Viana (1987) and Mata (1990) have described them prosodically. The first intonational unit is described with the contour rising-fall-rising, whereas the second unit exhibits a rising-fall contour. The prosody of tags is still studied too sparsely in EP (Mata (1990), for high school lectures, and Cruz-Ferreira (1998), for laboratory speech). For Cruz-Ferreira (1998), the tags are described with falling contours; while for Mata (1990) these structures are associated with rising ones. Furthermore, Falé (2005), Falé & Faria (2006) offer evidence for the categorical perception of intonational contrasts between statements and interrogatives, showing that the most striking clue associated with the perception of interrogatives is pitch range (the H% boundary tone has to be higher than 2 semitones), whereas declaratives are mostly perceived based on final falls in the stressed syllable. As for prosodic phrasing, Frota (2000); Viana et al. (2007) consider two different levels of phrasing equating both of them to the intonational phrase (IP): the major IP and the minor IP, in line with Ladd (2008). Both minor and major IPs are marked with breaks 3 and 4, respectively, and the diacritics - and % are used for boundary tones to represent the different strengths of the IP.

Focusing now on an automatic processing perspective, recovering punctuation marks and disfluencies are two relevant MDA (Metadata Annotation) tasks. The impact of the methods and of the linguistic information on structural metadata tasks has been discussed in the literature. Christensen, Gotoh, and Renals, S. (2001) report a generic HMM (Hidden Markov Model) framework that allows the combination of lexical and prosodic clues for recovering full stops, commas and question marks. A similar approach was also used by Liu et al. (2006) and Shriberg, Stolcke, Hakkani-Tür, and Tür (2000) for detecting sentence boundaries. Other recent studies have shown that the best performance for the punctuation task is achieved when prosodic, morphological and syntactic information are combined (Favre, Hakkani-Tür, & Shriberg, 2009; Liu et al., 2006; Ostendorf, et al., 2008).

Much of the features and the methods used for sentence-like unit detection may be applied in disfluency detection tasks. What is specific of the latter is that disfluencies have an idiosyncratic structure: *reparandum*, *interruption point*, *interregnum* and *repair* of fluency (Levelt, 1989; Nakatani & Hirschberg, 1994; Shriberg, 1994). The *reparandum* is the region to repair. The *interruption point* is the moment when the speaker stops his/her production to correct the linguistic material uttered. Ultimately, it is the frontier between disfluent and fluent speech. The *interregnum* is an optional part and may include silent pauses, filled pauses (*uh*, *um*) or explicit editing expressions (*I mean*, *no*). The *repair* is the corrected linguistic material. It is known that each of these regions has idiosyncratic acoustic properties that distinguish them from each other, inscribed in the edit signal theory (Hindle, 1983), meaning that speakers signal an upcoming repair to their listeners. Based on the edit signal theory, Nakatani and Hirschberg (1994) and Shriberg et al. (2000) used CARTs (Classification and Regression Trees, Breiman, Friedman, Olshen, & Stone, 1984) to identify different prosodic features of the interruption point. Heeman & Allen (1999) present a statistical language model including the identification of part-of-speech tags, discourse markers, speech repairs, and intonational phrases, achieving better performances by analysing those events simultaneously. Kim, Schwarm, and Ostendorf (2004) and Liu et al. (2006) used features based on previous studies and added language models to predict both prosodic and lexical features of sentence boundaries and disfluencies.

Our approach is a classification task, like Kim et al. (2004), aiming at discriminating between comma, full stop, question marks, and disfluencies. However, contrarily to Kim et al. (2004) and Liu et al. (2006), which target SU boundary detection and IP detection, this study discriminates three distinct punctuation marks as possible boundary events, since it is known for Portuguese that the impact of several linguistic features is different for each punctuation mark (Batista et al*.*, 2012a). This work aims at classifying metadata structures based on the following set of very informative features derived from the literature: pause at the boundary, pitch declination over the sentence, post-boundary pitch and energy resets, pre-boundary lengthening, word duration, silent pauses, filled pauses, and presence of fragments. With this work we hope to contribute to the discussion of what are language and domain dependent effects in structural metadata evaluation (Kolár, Liu, & Shriberg, 2009; Kolár & Liu, 2010; Ostendorf et al., 2008; Shriberg et al., 2009).

3 Targeted corpora

This work focuses on two domains to discriminate between structural metadata events, namely: university lectures and dialogues. The choice of the corpora was influenced by the availability of large amounts of (highly spontaneous) transcribed data in European Portuguese for these two domains. Since we are working on developing annotation techniques that can be applied to any corpus type, we started with those two domains representative of spontaneous data, a very hard challenge. If the annotation techniques would be suitable for spontaneous data, then they would be more easily extended to other domains. Moreover, both corpora are in the process of being publicly available through the META-NET infrastructure (http://metanet4u.eu/).

LECTRA is a university lectures corpus, collected within the LECTRA national project (Trancoso, Martins, Moniz, Mata, & Viana, 2008) and aiming at producing multimedia contents for e-learning applications, and also at enabling hearing-impaired students to have access to recorded lectures. It includes seven 1-semester courses, six of them recorded in the presence of students, and only one recorded in a quiet environment. Most classes are 60-90 minutes long, targeting an Internet audience. The initial set of 21 hours orthographically transcribed was recently extended to 32 hours (Pellegrini, Moniz, Batista, Trancoso, & Astudillo, 2012). The corpus was divided into 3 different sets:

train (78%), development (11%), and test (11%). The sets include portions of each one of the courses and follow a temporal criterion, meaning the first classes of each course were included in the training set, whereas the final ones were integrated into both development and test sets.

CORAL is a corpus of map-task dialogues (Anderson et al., 1991), collected by Trancoso, Viana, Duarte, and Matos (1998). One of the participants (giver) has a map with some landmarks and a route drawn between them; the other (follower) has also landmarks, but no route and consequently must reconstruct it. In order to elicit conversation, there are small differences between the two maps: one of the landmarks is duplicated in one map and single in the other; some landmarks are only present in one of the maps; and some are synonyms. The 32 speakers were divided into 8 quartets and in each quartet organized to take part in 8 dialogues, totalling 64 dialogues, which corresponds to 9 h (46 k words). In the scope of this work, 20% of the corpus was used for evaluation.

3.1 Corpora annotation

CORAL and LECTRA share a core annotation schema, comprising orthographic, morpho-syntactic, structural metadata, and paralinguistic information (laughs, coughs, etc.). The multi-layer annotation aimed at providing a suitable sample for further linguistic and speech processing analysis. A full report on this can be found in Trancoso et al. (2008) and Moniz (2013).

The orthographic manual transcriptions were produced based on initial automatic transcripts, which were manually corrected. Speech was segmented into chunks delimited by silent pauses, already containing audio segmentation related to speaker and gender identification and background conditions. All segmented speech chunks are manually punctuated following the punctuation descriptions detailed in Duarte (2000) and annotated. The annotation of disfluencies is provided in a separate tier, closely following Shriberg (1994) and basically using the same set of labels. Following an analysis of Eklund (2004), disfluent items are also indexed. Another level of manual annotation aimed both at providing basic syntactic information and a speech segmentation into sentence-like units (SUs), following the Simple Metadata Annotation Specification from the Linguistic Data Consortium. The corpora were also annotated with part-of-speech tags using *Falaposta* (Batista, Moniz, Trancoso, Mamede, & Mata, 2012b), a Conditional Random Field-based tagger that accounts for 28 part-of-speech tags and achieves 95.6% accuracy. *Falaposta* was also used to detect foreign words in both corpora. Foreign words in the lectures accounts for 0.4% (989), whereas in dialogues they are almost inexistent (just one occurrence).

3.2 Automatic transcription

The speech recognition system Audimus (Neto et al., 2008) was used to produce all the automatic speech transcripts used in the context of this work. Figure 1 depicts the Automatic Speech Recognition (ASR) offline-processing pipeline. The first module, after jingle detection, performs audio diarization, which consists of assigning the portions of speech to a speaker and also of classifying the gender of the speaker. Several modules are also integrated in the speech recognizer pipeline, such as capitalization and punctuation, topic segmentation and indexing, and summarization.

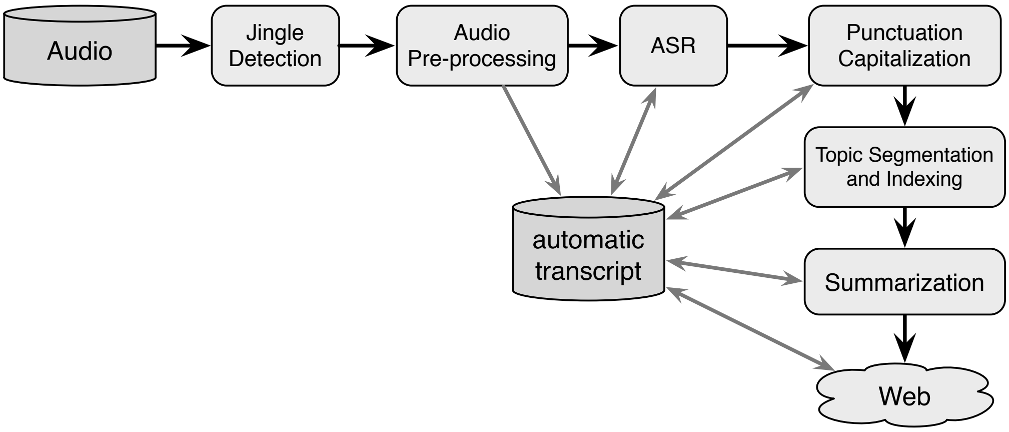


Figure 1. ASR offline-processing architecture

The ASR module is a hybrid automatic speech recognizer that combines the temporal modelling capabilities of Hidden Markov Models with the pattern discriminative classification capabilities of Multi-layer Perceptrons. The speech recognizer was trained for the broadcast news domain, being unsuitable for other domains, such as map-task dialogues or university lectures. Therefore, for a number of experiments, the ASR was used in a forced align mode, in order not to bias the study with the bad results obtained with an out-of-domain recognizer.

3.3 Manual/automatic transcript synchronization

Manual transcripts are usually composed of segments, containing information about their start and end locations in the signal file, not necessarily at the word level. Depending on the purpose, manual transcripts may also include a wide range of additional information for a given speech region, such as: speaker id, speaker gender, focus conditions, sections to be excluded from evaluation, segmentation information, punctuation marks, disfluency marking, capitalization, metadata indicating the presence of foreign languages, and other phenomena. Automatic transcripts, produced by ASR systems, differ from manual transcripts in many ways: ASR data usually corresponds to a sequence of lowercase words, each of which referring its corresponding time period in the speech signals and its confidence score. Besides words, automatic speech transcripts may also include additional information that can be extracted directly from the audio signal, such as background speech conditions (clean/noise/music), speaker identification, speaker gender, phone information, and other metadata.

The experiments performed in this work are based on automatic transcripts, produced in forced alignment mode. However, the corresponding manual transcripts also provide complementary reference data that are fundamental for speech analysis, supervised training, and automatic evaluation. Instead of creating task dependent links between the ASR output and the corresponding manual transcripts, the relevant manual annotations, as well as other available elements, were automatically transferred into the automatic transcript itself, in a way that small programs can easily process it. Such a self-contained dataset, merging all the relevant information, can be easily dealt with and constitutes a valuable resource for extensively address, study and process speech data. Therefore, the first stage in our framework consists of extending existing ASR transcripts in order to accommodate relevant reference data coming from manual annotations.

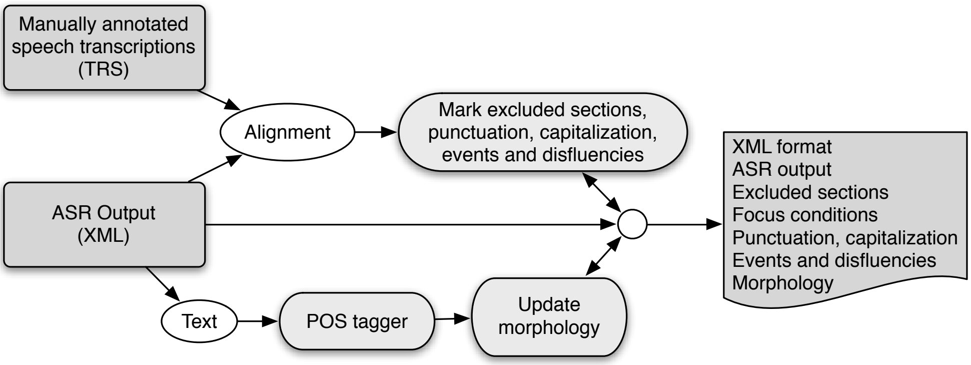
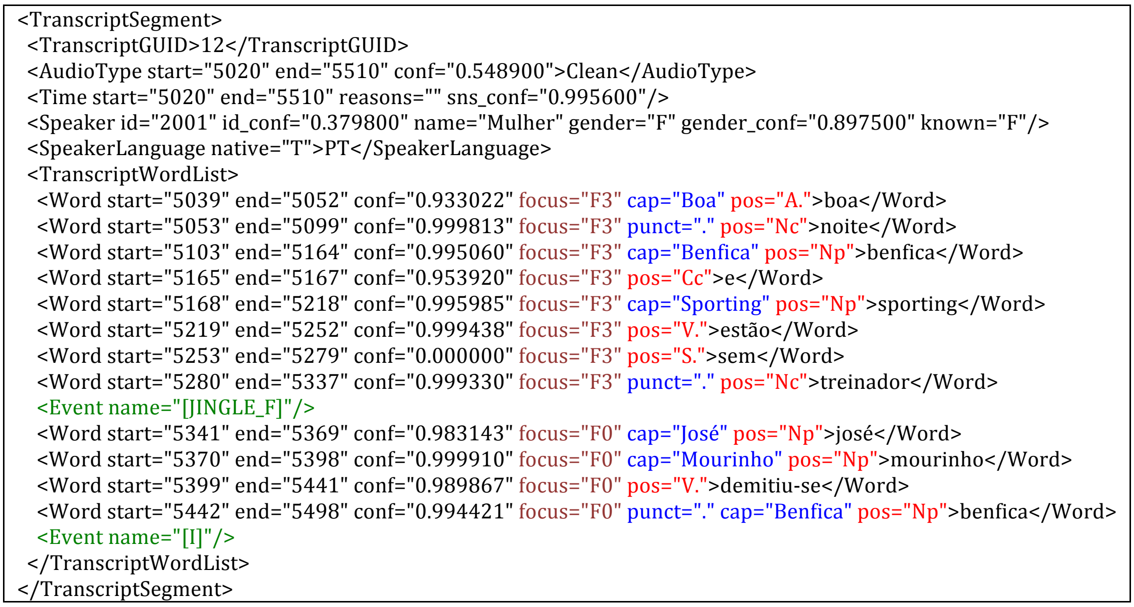

Manually annotated speech transcriptions (TRS)
Alignment
Mark excluded sections, punctuation, capitalization, events and disfluencies
ASR Output (XML)
XML format
ASR output
Excluded sections
Focus conditions
Punctuation, capitalization
Events and disfluencies
Morphology
Text
POS tagger
Update morphology

Figure 2. Creating a file that integrates the reference data into the ASR output

The manual orthographic transcripts include punctuation marks, disfluency annotations, and capitalization information, which constitute our reference data. However, that is not the case of the fully automatic and forced align transcripts, which only include information such as: word time intervals and confidence scores. The required reference is provided by means of alignments between the manual and automatic transcripts, a non-trivial task due to recognition errors. Figure 2 illustrates this process, revealing also recent efforts on including paralinguistic information as well (breath pausing, laughs, etc.).



```
<TranscriptSegment>
  <TranscriptGUID>12</TranscriptGUID>
  <AudioType start="5020" end="5510" conf="0.548900">Clean</AudioType>
  <Time start="5020" end="5510" reasons="" sns_conf="0.995600"/>
  <Speaker id="2001" id_conf="0.379800" name="Mulher" gender="F" gender_conf="0.897500" known="F"/>
  <SpeakerLanguage native="T">PT</SpeakerLanguage>
  <TranscriptWordList>
    <Word start="5039" end="5052" conf="0.933022" focus="F3" cap="Boa" pos="A.">boa</Word>
    <Word start="5053" end="5099" conf="0.999813" focus="F3" punct="." pos="Nc">noite</Word>
    <Word start="5103" end="5164" conf="0.995060" focus="F3" cap="Benfica" pos="Np">benfica</Word>
    <Word start="5165" end="5167" conf="0.953920" focus="F3" pos="Cc">e</Word>
    <Word start="5168" end="5218" conf="0.995985" focus="F3" cap="Sporting" pos="Np">sporting</Word>
    <Word start="5219" end="5252" conf="0.999438" focus="F3" pos="V.">estão</Word>
    <Word start="5253" end="5279" conf="0.000000" focus="F3" pos="S.">sem</Word>
    <Word start="5280" end="5337" conf="0.999330" focus="F3" punct="." pos="Nc">treinador</Word>
    <Event name="[JINGLE_F]"/>
    <Word start="5341" end="5369" conf="0.983143" focus="F0" cap="José" pos="Np">josé</Word>
    <Word start="5370" end="5398" conf="0.999910" focus="F0" cap="Mourinho" pos="Np">mourinho</Word>
    <Word start="5399" end="5441" conf="0.989867" focus="F0" pos="V.">demitiu-se</Word>
    <Word start="5442" end="5498" conf="0.994421" focus="F0" punct="." cap="Benfica" pos="Np">benfica</Word>
    <Event name="[I]"/>
  </TranscriptWordList>
</TranscriptSegment>
```

Figure 3. Creating a file that integrates the reference data into the ASR output

The resulting file corresponds to the ASR output, extended with: time intervals to be ignored in scoring, focus conditions, speaker information for each region, punctuation marks, capitalisation, and disfluency annotation. Figure 3 shows an automatic transcript segment, enriched with reference data, corresponding to the sentence: *Boa noite. Benfica e Sporting estão sem treinador. José Mourinho demitiu-se [do] Benfica* ‘Good evening. Benfica and Sporting have no coach. José Mourinho resigned from Benfica’.

The example illustrates two important sections: the characterization of the transcript segment and the discrimination of the wordlist that comprises it. Each word element contains the lowercase orthographic form, start time, end time, and confidence level; a discrimination of the focus condition (*F3* stands for speech with music and *F0* stands for planned speech without background noise); information about the capitalized form (*cap*); whether or not it is followed by a punctuation mark (*punct*=.); and the part-of-speech tag (*pos=A*, where *A* stands for adjective).

4 Integrating prosodic information

The extended self-contained automatic transcripts, described in the previous section, serves as a good data source for a number of experiments that rely purely on lexical and audio segmentation features. However, the integration of the multi-layer grammatical knowledge must also account for prosodic information. The remainder section describes the prosodic feature extraction process and the creation of an improved data source, containing additional prosodic information, aiming at an organizational structure of speech.

4.1 Phone and pause duration

In addition to monophone units modelled by a single state, the ASR system uses multiple-state monophone units, and a fixed set of phone transition units, generally known as diphones, aimed at specifically modelling the most frequent intra-word phone transitions (Abad & Neto, 2008). An analysis of 1 h of manually transcribed speech revealed several problems in the boundaries of silent pauses, and in their frequent misdetection, which affected the phone boundaries. Based on such analysis, we have designed a tool for converting the phones/diphones information into monophones. Figure 4 presents an excerpt of the resulting information. Alternatively, the speech recognition could be performed with monophones, but the Word Error Rate would increase.

```
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.000 0.270 interword-pause
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.270 0.040 "m
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.310 0.040 u~
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.350 0.035 j~
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.385 0.035 #t
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.420 0.030 u+
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.450 0.040 "b
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.490 0.120 o~+
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.610 0.070 "d
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.680 0.075 i
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.755 0.025 #A+
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.780 0.060 interword-pause
```

Figure 4. Monophones marked with syllable boundary (the diacritic #) and stress (the diacritic ")

4.2 Marking the syllable boundaries and stress

Another important step consisted of marking the syllable boundaries as well as the syllable stress, tasks that were absent in the recognizer. The extensive literature on prosodic studies (e.g., Beckman & Pierrehumbert, 1986; Frota, 2000 and 2009; Gussenhoven, 2004; Ladd, 2008; Nespor & Vogel, 2007; Pierrehumbert, 1980; Pierrehumbert & Hirschberg, 1990; Viana, 1987) suggests that tonic and post-tonic syllables are of crucial importance to account for different prosodic aspects, such as, nuclear and boundary tones, duration of those units, or even rhythmic patterns. The task of marking syllable boundaries and stress was achieved by means of a lexicon containing all the pronunciations of each word together with syllable information. A set of syllabification rules was designed and applied to the lexicon. The rules account fairly well for the canonical pronunciation of native words, but they still need improvement for words of foreign origin. The excerpt presented in Figure 4 shows an example of marked syllable boundaries and stress.

4.3 Extracting pitch and energy

By the time these experiments were conducted, pitch (f_0) and energy (E) information were not available in the ASR output. For that reason, it has been directly extracted using the Snack toolkit (Sjölander et al., 1998). Algorithms for automatic extraction of the pitch track have, however, some problems, e.g., octave jumps; irregular values for regions with low pitch values; disturbances in areas with micro-prosodic effects; influences from background noisy conditions; inter alia. Several tasks were needed in order to solve some of these issues. We have removed all the pitch values calculated for unvoiced regions in order to avoid constant micro-prosodic effects. This is performed in a phone-based analysis by detecting all the unvoiced phones. As to the influences from noisy conditions, the recognizer classifies the input speech according to different focus conditions (e.g., noisy, clean), making it possible to isolate speech segments with unreliable pitch values. Figure 5 illustrates the process described above, where the original pitch values are represented by dots and the grey line represents the resultant pitch. Due to this process, a significant number of pitch values were changed in both corpora (13.6% for Lectra, and 5.6% for Coral). The first tier is the orthographic tier, also containing part-of-speech tags; the second tier corresponds to the multiple-state monophone/diphone units, and the last tier is the resulting conversion for monophones.

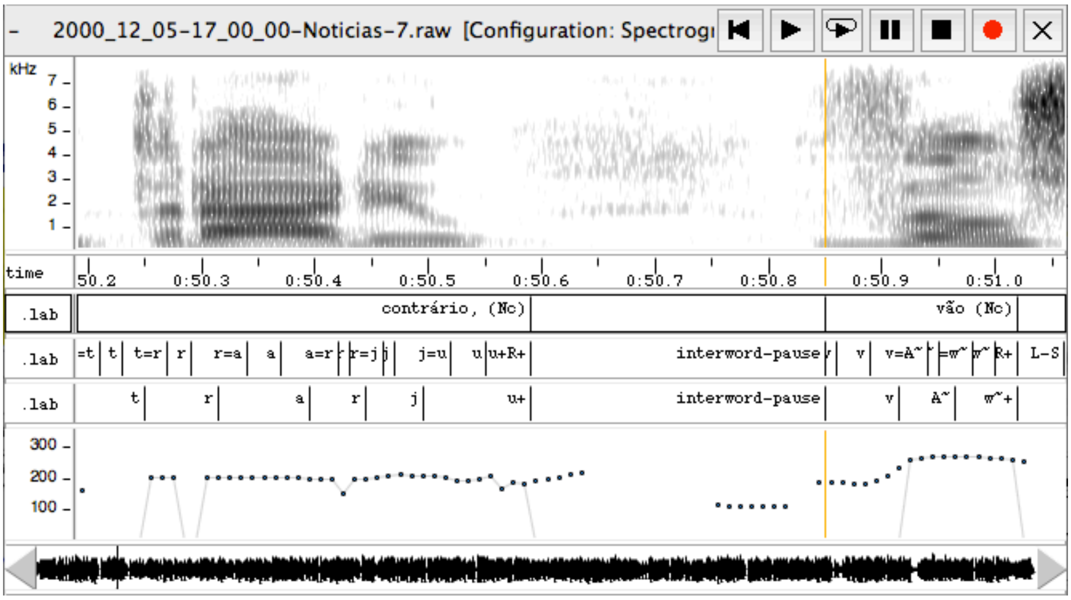

2000_12_05-17_00_00-Noticias-7.raw [Configuration: Spectrogı
kHz
7
6
5
4
3
2
1
time
50.2
0:50.3
0:50.4
0:50.5
0:50.6
0:50.7
0:50.8
0:50.9
0:51.0
.lab
contrário, (Nc)
vão (Nc)
.lab
=t t t=r r r=a a a=r r=j j j=u u u+R+ interword-pause v v=A~ =w~ w~ R+ L-S
.lab
t r a r j u+ interword-pause v A~ w~+
300
200
100

Figure 5. Pitch adjustment

4.4 *Producing the final transcript*

After extracting and calculating the above information, the existing data source was upgraded in order to accommodate the additional prosodic information. Figure 6 illustrates the involved processing steps, where PCTM denotes a *phone-level time-marked conversation file*. Pitch and energy values are extracted directly from the speech signal. A Gaussian Mixture Model classifier is then used to automatically detect speech/non-speech regions, based on the energy. Both pitch and speech/non-speech values are then used to adjust the boundaries of the acoustic phone transitions.

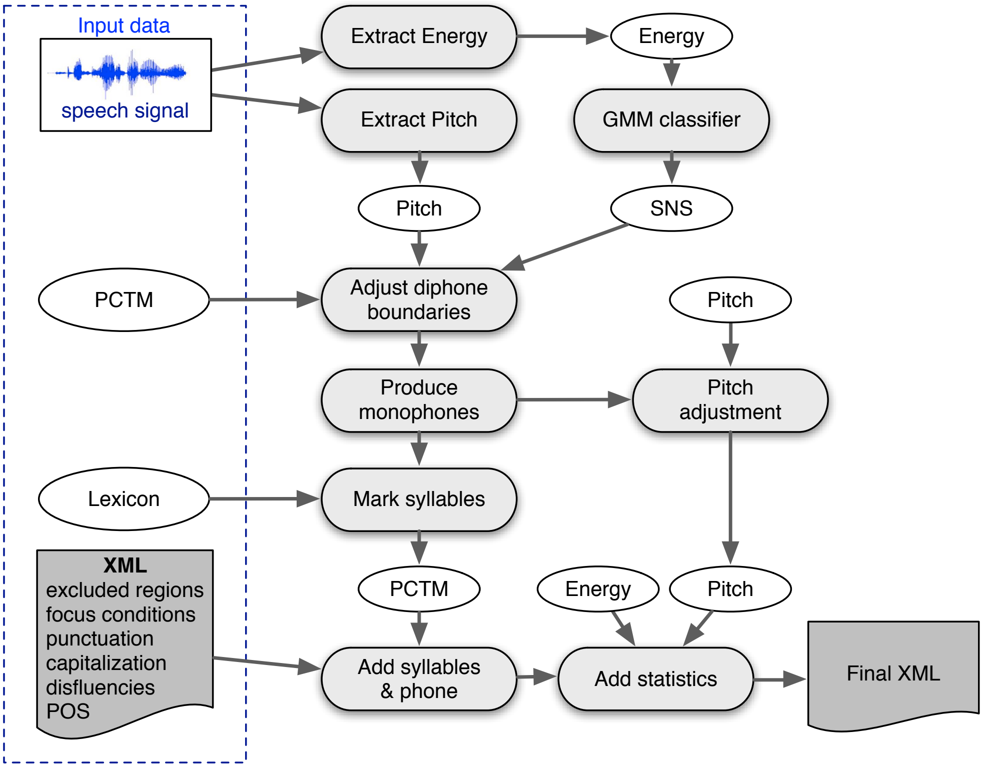


Figure 6. Integrating prosodic information.

```
<Word start="5053" end="5099" conf="0.999813" focus="F3" punct="." pos="Nc" name="noite"
        phseq="_noj#t@+" pmax="268.3" pmin="74.8" pavg="212.3" pmed="209.1" pstdev="34.57" emax="62.8"
        emin="32.9" eavg="52.1" emed="58.2" estdev="10.17" eslope="-0.3" eslope_norm="-12.14" pslope="-
        0.18" pslope_norm="-8.41" pmin_st_100="-5.03" pmax_st_100="17.09" pmin_st_spkr="1.70"
        pmax_st_spkr="23.81" pmin-zscore_spkr="-3.25">
 <syl stress="y" start="5053" dur="25.5" pmax="268.3" pmin="196.2" pavg="218.0" pmed="210.2"
        pstdev="20.60" emax="62.8" emin="37.9" eavg="56.9" emed="59.3" estdev="5.98" eslope="-0.2"
        eslope_norm="-4.89" pslope="0.17" pslope_norm="4.08">
  <ph name="n" start="5053" dur="7" pmax="216.2" pmin="201.0" pavg="208.5" pmed="209.1" pstdev="4.41"
        emax="60.1" emin="52.9" eavg="55.7" emed="54.6" estdev="2.78"/>
  <ph name="o" start="5060" dur="9" pmax="215.3" pmin="196.2" pavg="203.0" pmed="200.1" pstdev="6.37"
        emax="60.5" emin="58.5" eavg="59.5" emed="59.5" estdev="0.63"/>
  <ph name="j" start="5069" dur="9.5" pmax="268.3" pmin="221.2" pavg="243.3" pmed="241.1"
        pstdev="15.49" emax="62.8" emin="37.9" eavg="55.4" emed="60.3" estdev="8.81"/>
 </syl>
 <syl start="5078.5" dur="21.5" pmax="74.8" pmin="74.8" pavg="74.8" pmed="74.8" pstdev="0.00" emax="60.9"
        emin="32.9" eavg="45.9" emed="40.6" estdev="11.03" eslope="1.1" eslope_norm="23.29" pslope="0.00"
        pslope_norm="0.00">
  <ph name="t" start="5078.5" dur="8.5" emax="40.0" emin="32.9" eavg="35.8" emed="35.1" estdev="2.22"/>
  <ph name="@" start="5087" dur="13" pmax="74.8" pmin="74.8" pavg="74.8" pmed="74.8" pstdev="0.00"
        emax="60.9" emin="33.2" eavg="52.8" emed="58.0" estdev="9.12"/>
 </syl>
</Word>
```

Figure 7. Excerpt of the final XML, containing information about the word *noite*‘*night’*

The final XML file combines all the previous information together with pitch and energy statistics for each unit. Figure 7 shows an excerpt of the resultant data, containing the information about a word. *syl* stands for syllable, *ph* for phone, *p** means pitch, *e** energy, and *dur* corresponds to a duration (measured in 10ms frames). Information concerning words, syllables and phones can be found in the file, together with pitch, energy and duration information. For each unit of analysis we have calculated the minimum, maximum, average, and median, for both pitch and energy raw and normalised values. Pitch slopes were also calculated after converting the pitch into semitone values.

4.5 Prosodic features

Our experiments use a fixed set of purely automatic features, extracted or calculated from the extended transcripts, described in the previous subsection. The features involve the word itself and the adjacent words. Features involving a single word include: pitch and energy slopes; ASR confidence score; word duration; number of syllables and number of phones. Features involving two consecutive words include: pitch and energy slopes shapes; pitch and energy differences; comparison of durations and silences before each word; and ratios for silences, word durations, pitch medians, and energy medians. Pitch slopes were calculated based on semitones rather than frequency. Slopes in general were calculated using linear regression. Silence and duration comparisons assume 3 possible values, expanding to 3 binary features: > (greater than), = (equal), or < (less than). The ratios assume values between 0 and 1, indicating whether the second value is greater than the first. The pitch and energy shapes are based on slope values and expand to 9 binary features, assuming one of the following values {RR, R-, RF, -R, --, -F, FR, F-, FF}, where F = Fall (if the slope is negative), - = plateau (if the slope is near zero), R = Rise (otherwise).

5 Discriminating between structural metadata events

The availability of prosodic annotated data allows us to simultaneously study sentence-like units and disfluencies, a step forward in structural metadata events analysis, since our previous studies did not account for a combined approach of all those events and focus on audio segmentation and lexical features.

The main motivations for studying the prosodic properties of both disfluencies and punctuation marks are three: (i) the location of a disfluency may be automatically and manually misclassified with the one of a punctuation mark, i.e., they may share prosodic properties; (ii) for speech recognition, in particular, and for linguistics, in general, the analysis of the prosodic patterns of a disfluency are important to a better understanding of phrasing and intonational contours; and (iii) studies conducted in European Portuguese concentrate on small data, mostly laboratory data, therefore, a data-driven approach of the prosodic patterns extracted from structural metadata would be a contribution to the understanding of both speech acts and (dis)fluent patterns in European Portuguese.

This section describes the most recent results in discriminating between structural metadata events and regular words.

5.1 Most recent results

Table 1 displays the structural metadata events of both university lectures and dialogues. This table shows some differences (namely in terms of word per minute and punctuation marks) and some similarities. Contrarily to what would be expected, the map-task corpus does not elicit more disfluencies in average per minute. In the dialogues, sentences have fewer words and are shorter than the sentences produced by a teacher (128 words *vs.* 79 words per minute). The number of words per minute in the lectures (128 words) compares well with the 120 words pointed out by Aiken, Thomas, and Shennum (1975) as a normal speech rate in a lecture. The number of words per minute in the dialogues, although a raw measure not accounting for long or short turns as in Burger and Sloane (2004), is quite inferior to lectures, as expected. The total number of turns in dialogues is 5,549, with a minimum of a single word up to a maximum of over 50 words per turn (very rare cases).

Table 1. Lectra and Coral properties

Corpora →	**Lectra**	**Coral**		
time (h)	31:24	9:42	Per minute	
number of words + filled pauses	240,951	45,884	128	79
number of disfluent sequences	9,340	2,248	5	4
disfluencies followed by a repair	6,328	1,531	3	3
number of full stops	9,224	6,000	5	10
number of commas	25,569	4,552	14	8
number of question marks	4,024	1,179	2	2

For the sake of comparison, in other domains such as broadcast news, which typically contains more read than spontaneous speech, the number of words per minute (161 words; Batista et al., 2012a) is higher than in lectures and dialogues.

Reasoning about the main differences, the on-the-fly editing process in a map-task dialogue implies a straight cooperative process between two interlocutors. The workflow of the dialogue is made of mutual contributions of turns, silences, and a plethora of paralinguistic events established in separate turns or in overlapping ones. The dialogic nature is ultimately about active speaking and active listening as a joint process (Campbell, 2010). In lectures, although informal and spontaneous, there is a main speaker holding the floor and allowing for interactions, but by far not as frequent as in dialogues.

One important aspect that characterizes Portuguese punctuation marks is the high frequency of commas. In a previous study Batista et al. (2012a), where Portuguese and English broadcast news are compared, the percentage of commas in the former is twice the frequency of the latter. Those findings still stand for the university lectures corpus, accounting for more than 50% of all events, but not for the dialogues, where the full stops are even more frequent than commas. Moreover, it is also interesting to note that both corpora elicit around 2 questions per minute. Most of the questions in dialogues is made by the follower seeking for clarifications; whereas the ones in lectures are very frequently tag questions, so the auditory can acknowledge what was said by the teacher, or even rhetorical questions.

Our experiments applied distinct statistical methods: Naïve Bayes, Logistic Regression, J48, and CART. The best results were consistently achieved using CART. Results were evaluated using the standard performance metrics (Makhoul, Kubala, Schwartz, & Weischedel, 1999): *Precision*, *Recall*, *F-measure* and *SER* (Slot Error Rate), where each slot corresponds to a metadata event.

$$Precision = \frac{correct}{M}, Recall = \frac{correct}{N}, Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}, SER = \frac{errors}{N}$$

where *N* is the number of events in the reference, *M* is the number of events in the hypothesis, *correct* are the number of correct events, and *errors* corresponds to the number of event errors (misses and false alarms). For example, for the task of detecting structural elements, the SER corresponds to the number of incorrectly classified elements over the number of structural elements in the reference data.

Table 2. CART classification results for prosodic features

	Lectra				Coral			
Class	Precision	Recall	F	SER	Precision	Recall	F	SER
comma (,)	60.6	27.6	37.9	90.3	57.1	38.5	46.0	90.5
full stop (.)	64.1	67.6	65.8	70.2	65.4	66.8	66.1	68.5
question (?)	73.9	29.5	42.2	80.9	48.4	19.7	28.0	101.3
repair	60.8	13.1	21.6	95.4	72.4	27.7	40.1	82.9
weighted avg.	63.0	32.9	43.3	75.6	62.5	47.9	54.2	61.7

Our experiments consider four different classes of structural elements, full stops, commas, question marks, and disfluency repairs. We aim at automatically detecting structural metadata events and at discriminating between them, using mostly prosodic features (with the exception of the feature *two identical contiguous words*). Table 2 presents the best results achieved. For both corpora the best performance is achieved for full stops, confirming our expectation, since prosodic information is known to be crucial to classify those events in Portuguese. The low results concerning commas are also justifiable, because our experiments rely on prosodic features, but commas depend mostly on lexical and syntactic features (Favre et al., 2009). The performance for question marks is mainly related to their lower frequency and to the multiple prosodic patterns found for these structures. Interrogatives in Portuguese are not commonly produced with subject-auxiliary verb inversion, as in English, which renders the problem of identifying interrogatives even more challenging. Moreover, in dialogues the most frequent interrogatives are elliptic interrogatives uttered as a single *Sim*? ‘Yes?’, making the task even harder. The worst performance in university lectures, especially affected by a low recall, is achieved for repairs. While prosodic features seem to be strong cues for detecting this class, the confusion matrix presented in Table 3 reveals that repairs are still confused with regular words.

Table 3. Confusion matrix between events

Lectra

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

Coral

classified as →	,	.	?	rep.	del
comma (,)	**347**	108	5	17	425
full stop (.)	100	**762**	34	4	241
question (?)	12	137	**46**	0	39
repair	41	37	3	**89**	151
insertions	108	121	7	13	

Our recent experiments as well as other reported work (Levelt, 1989; Nakatani & Hirschberg, 1994; Shriberg, 1994) suggest that filled pauses and fragments serve as strong cues for detecting structural regions of a disfluent sequence. Supported by such facts, we have conducted additional experiments using filled pauses and fragments as features. These turned out to be amongst the most informative features, highly increasing the repair *F-measure* in university lectures (21.6% → 48.8%) and dialogues (40.1% → 59%); and improving the overall *F-measure* to 47.8% and 57.2%, respectively. However, the impact of fragments is lower than the one reported by Nakatani and Hirschberg (1994) and Kim et al. (2004), due to the fact that fragments represent only 6.6% of all disfluent types in the lectures and 14.9% in the dialogues. As a final note, we would like to mention that our results use mostly language independent prosodic features, which explains the smaller performance when comparing to other reported state-of-the-art work, such as Batista et al. (2012a), Liu et al. (2006), and Ostendorf et al. (2008).

5.2 Most salient features

A set of informative features stands out as determinant to disambiguate between metadata events, namely, pitch and energy shapes, duration ratios, and confidence levels of the units of analysis. However, those features have distinct weights accordingly to the specific corpora. The most striking difference is the fact that lectures exhibit the highest pitch values per sentence, whereas dialogues exhibit the highest energy values, being those features significant different at $p < 0.001$ ($z = -23.598$ for pitch median and $z = -25.325$ for pitch slopes; $z=-20.599$ for energy median and $z=-14.849$ for energy slopes). Consequently, in lectures, pitch shapes and slopes are more discriminant for classifying structural metadata events, whereas in dialogues, energy features have a higher impact.

Another difference to consider in the features' impact in both corpora is the fact that full stops and question marks are better discriminated in lectures than in dialogues. As described in the previous section, question marks are scarcely classified in dialogues due mostly to the high frequency of elliptic questions composed of a single *Sim?* 'Yes'. The feature with higher impact in the dialogues is the *word.confidence.level*.

Relevant features for identifying the repair in both corpora comprise: (i) two identical contiguous words; (ii) both energy and pitch increases in the following word and (mostly) a plateau contour on the preceding word; and (iii) a higher confidence level for the following word than for the previous word. This set of features reveals that repetitions are being identified, that repair regions are characterized by prosodic contrast marking (increases in pitch and energy) between disfluency-fluency repair (as in Moniz et al., 2012 and Moniz, 2013), and also that the first word of the repair has a higher confidence score. Since repetitions are more frequent in dialogues (22% of all disfluencies *vs.* 16% in lectures), the feature *identical.contiguous.words* has a significant higher impact.

As for full stops, the determinant prosodic features correspond to: (i) a falling contour in the previous word; (ii) a plateau energy slope in the previous word; (iii) the duration ratio between the previous and the following words; and (iv) previous word higher confidence score. This characterization is the one that most resemble the neutral statements in Portuguese, with the canonical contour H+L* L% (Frota, 2000), associated with terminus value.

Question marks in lectures are characterized by two main patterns: (i) a rising contour in the current word and a rising/rising energy slope between previous and following words; and (ii) a plateau pitch contour in the previous word and a falling energy slope in the previous word. The rising patterns associated with question marks are not surprising, since they are commonly associated with interrogatives, since interrogatives are cross-language perceived as having a rising contour (Hirst & Di Cristo, 1998). The falling pitch contours have also been ascribed for different types of interrogatives, especially wh- questions in Portuguese.

Commas are the event characterized by fewest prosodic features, being mostly identified by morpho-syntactic features. However, that is not the case in dialogues, where they are better classified. The two most relevant features are: *identical.contiguous.words* and mostly plateau energy and pitch shapes between words. The first feature is associated with emphatic repetitions, comprising several structures, namely: (i) affirmative or negative back-channels (*sim, sim, sim* ‘yes, yes, yes’) and (ii) repetition of a syntactic phrase, such as a locative prepositional phrase (*para cima, para cima* ‘above, above’). They are used for precise tuning with the follower and for stressing the most important part of the instruction. Emphatic repetitions are annotated with commas separating the repeated item(s) and account for 1% of the total number of words in dialogues. Although not a disfluent structure, if they were accounted as disfluent words, they would represent 16.7% of all disfluent items. As for the features energy and pitch plateau shapes between words they are linked to lists of enumerated names of the landmarks in a given map, at the end of a dialogue.

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

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

6 Conclusions and future work

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

Future experiments will extend this study to fully ASR transcripts and evaluate how the discrimination between the punctuation marks and disfluencies is affected by the ASR errors. Future work will also tackle the inclusion of lexical and morpho-syntactic features, which are expected to considerably improve the performance, especially for commas and question marks. Future work will also include in the created framework very recent encouraging experiments on automatic labelling of ToBI adapted to European Portuguese, adding more discriminative prosodic information beyond pitch shapes and slopes, aiming at a systematization of intonational contours for several domains.

The work presented in this chapter is quite flexible and the framework is currently being used with other languages. In order to do so, the authors are currently working on identifying pseudo-syllables rather than using language dependent sets of rules. To the best of our knowledge, this is the first work conducted for European Portuguese on annotation techniques for ASR outputs and prosodic labelling, inscribing our project in very recent multidisciplinary projects in the Romance Languages, such as the one developed by Garrido et al. (2013) for Spanish and Catalan or by Mertens (2004) for French.

Acknowledgements

Helena Moniz is supported by FCT grant SFRH/BPD/95849/2013. Fernando Batista is supported by DCTI - ISCTE-IUL. This work was partially funded by PTDC/CLE-LIN/120017/2010 COPAS, and by PEst-OE/EEI/LA0021/2013. The authors would like to thank the anonymous reviewers for their very helpful comments and suggestions.

References

Abad, A., & Neto, J. (2008). Incorporating acoustical modelling of phone transitions in a hybrid ANN/HMM speech recognizer. In *Proceedings of Interspeech 2008*, Brisbane, Australia, 2394-2397.

Aiken, E., Thomas, G., & Shennum, W. (1975). Memory for a lecture: effects of notes, lecture rate, and informational density. *Journal of Educational Psychology*, *67*(3), 439-444.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34*, 351-366.

Batista, F., Moniz, H., Trancoso, I., & Mamede, N. (2012a). Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts. *IEEE Transactions on Audio, Speech, and Language Processing*, *20* (2), 474 – 485.

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

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook III*, 15-70.

Beckman, M., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Eds.) *Prosodic typology. The phonology of intonation and phrasing* (pp. 9–54). Oxford: Oxford University Press.

Bolinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse*. London: Arnold.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Taylor and Francis Group, New York.

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: Gleerup.

Burger, S., & Sloane, Z. (2004). The ISL Meeting Corpus: Categorical Features of Communicative Group Interactions. *Proceedings of Rich Transcription 2004, Spring Meeting Recognition Workshop,* Montreal, Canada.

Campbell, N. (2010). Expressive Speech Processing and Prosody Engineering: An Illustrated Essay on the Fragmented Nature of Real Interactive Speech. In F. Chen, & Jokinen, K. (Eds.), *Speech Technology Theory and Applications* (pp. 105 – 120). New York: Springer.

Christensen, H., Gotoh, Y., & Renals, S. (2001). Punctuation annotation using statistical prosody models. In *Proceedings of ASRU 2001*, Madonna di Campiglio, Italy, 35–40.

Cruz-Ferreira, M. (1998). Intonation in European Portuguese. In D. Hirst, & A. Di Cristo, (Eds.), *Intonation systems* (pp 167–178). Cambridge: Cambridge University Press.

Duarte, I. (2000). *Língua Portuguesa - Instrumentos de análise*. Universidade Aberta.

Eklund, R. (2004). *Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues*. PhD thesis, University of Linköpink.

Falé, I. (2005). *Percepção e reconhecimento da informação entoacional em Português Europeu*. PhD thesis, University of Lisbon.

Falé, I., & Faria, I. (2006). Categorical perception of intonational contrasts in European Portuguese. In *Speech Prosody 2006*, Dresden, Germany.

Favre, B., Hakkani-Tür, D., & Shriberg, E. (2009). Syntactically-informed Models for Comma Prediction. In *Proceedings of ICASSP'09*, Taipei, Taiwan.

Frota, S. (2000). *Prosody and Focus in European Portuguese*. *Phonological Phrasing and Intonation*. New York: Garland Publishing.

Frota, S. (2002). Nuclear falls and rises in European Portuguese: a phonological analysis of declarative and question intonation. In *Probus*, *14*, 113–146.

Frota, S (2009). The intonational phonology of European Portuguese. In S.-A. Jun (Ed.) *Prosodic typology – the phonology of intonation and phrasing.*Oxford*:* Oxford University Press, 6-42.

Frota, S. (2012). Prosodic structure, constituents and their representations. In A. Cohn, C. Fougeron, & M. Huffman (Eds.), *The Oxford Handbook of Laboratory Phonology*. Oxford: Oxford University Press, 255-265.

Garrido, J., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., Mota, C., González, C., Vivaracho, C., Rustullet, S., Larrea, O., Laplaza, Y., Vizcaíno, F., Estebas, E., Cabrera, M., & Bonafonte, A. (2013). Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Journal of Language Resources and Evaluation*, 47, 945-971.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

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

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. In *Proceedings of ACL-83*, 123-128.

Hirst, D., & A. Di Cristo (1998). *Intonation systems*. Cambridge: Cambridge University Press.

Jun, S.-A. (2005). *Prosodic typology – the phonology of intonation and phrasing.* Oxford University Press.

Kim, J., Schwarm, S.E., & Ostendorf, M. (2004). Detecting structural metadata with decision trees and transformation-based learning. In *Proceedings of HLT-NAACL 2004*, New York, U.S.A., 137–144.

Kolár, J., Liu Y., & Shriberg, E. (2009). Genre effects on automatic sentence segmentation of speech: a comparison of broadcast news and broadcast conversations. In *Proceedings of ICASSP 2009,* 4701-4704.

Kolár, J., & Liu, Y. (2010). Automatic sentence boundary detection in conversational speech: a cross-lingual evaluation on English and Czech. In *Proceedings. of ICASSP 2010*, 5258-5261.

Ladd, R. (2008). *Intonation Phonology*. Cambridge:Cambridge University Press.

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

Liberman, M. (1975). *The intonational system of English*. PhD Dissertation, MIT.

Liu, Y., Shriberg, E., Stolcke, A., Dustin, H., Ostendorf, M., & 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., & Weischedel, R. (1999). Performance measures for information extraction. In *Proceedings of the DARPA Broadcast News Workshop*, Herndon, VA, 249-252.

Mata, A.I. (1990). *Questões de entoação e interrogação no Português. Isso é uma pergunta?* Master thesis, University of Lisbon.

Mertens, P. (2004). The Prosogram: semi-automatic transcription of prosody based on a tonal perception model. Speech Prosody 2004, Nara, Japan, 23-26.

Moniz, H., Batista, F., Trancoso, I., & Mata, A.I. (2012). Prosodic context-based analysis of disfluencies. In *Proceedings of Interspeech 2012*, Portland, U.S.A.

Moniz, H. (2013). *Processing disfluencies in European Portuguese*. PhD Thesis. University of Lisbon.

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

Nespor, M., & Vogel, I. (2007). *Prosodic Phonology*. Berlin/New York: Mouton de Gruyter.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., & Caseiro, D. (2008). Broadcast news subtitling system in Portuguese. *Proceedings of ICASSP'08*, 1561–1564.

Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tür, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J., Liu, Y., Makey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W., & Wooters, C. (2008). Speech Segmentation and Spoken Document Processing. *IEEE Signal Processing Magazine*, 25, 59-69.

Pellegrini, T., Moniz, H., Batista, F., Trancoso, I., & Astudillo, R. (2012). Extension of the LECTRA corpus: classroom LECture TRAnscriptions in European Portuguese. In Proceedings of *Speech and Corpora*, Belo Horizonte, 98-102.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. PhD thesis, MIT.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. M. & M. Pollack (Eds.), *Intentions in communication* (Press, pp. 271-311). Cambridge, MA: MIT.

Rosenberg, A. (2009). *Automatic detection and classification of prosodic events*. PhD thesis, Columbia University.

Shattuck-Hufnagel, S., & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *In Journal of Psycholinguistic Research*, *25*(2), 193-247.

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

Shriberg, E. (1999). Phonetic consequences of speech disfluency. In Proceedings of *ICPhS'99*, San Francisco, U.S.A., 612–622.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., & Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, *32*(1-2), 127–154.

Shriberg, E., Favre, B., Fung, J., Hakkani-Tür, D., & Cuendet, S. 2009. Prosodic similarities of dialog act boundaries across speaking styles. In S.-C. Tseng (Ed.), *Linguistic Patterns in Spontaneous Speech*, *Language and Linguistics Monograph Series* (pp. 213-239). Taipei: Institute of Linguistics, Academia Sinica.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: a standard for labeling English prosody. *In Proceedings of CSLP'98*, Banff, Canada, 867-870.

Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., & Granström, B. (1998). Web-based educational tools for speech technology. In *Proceedings of ICSLP* 1998, Sydney, Australia, 3217–3220.

Trancoso, I., Martins, R., Moniz, H., Mata, A. I., & Viana, M.C. (2008). The LECTRA Corpus - Classroom Lecture Transcriptions in European Portuguese. In *Proceedings LREC'08*, Marrakech, Morocco.

Trancoso, I., Viana, M. C., Duarte, I., & Matos, G. (1998). Corpus de Diálogo CORAL, In *Proceedings of PROPOR'98*, Porto Alegre, Brasil.

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler,. &R. Ladd (Eds.), *Prosody: models and measurements* (pp. 55–66). Berlin: Springer.

Viana, M.C. (1987). *Para a Síntese da Entoação do Português*. PhD thesis, University of Lisbon.

Viana, M.C., Frota, S., Falé, I., Mascarenhas, I., Mata, A. I., Moniz, H., & Vigário, M. (2007). Towards a P_ToBI. In *Proceedings of PaPI 2007*, Minho, Portugal.

Vigário, M. (1995). *Aspectos da prosódia do Português Europeu. Estruturas com advérbios de exclusão e negação frásica*. Master thesis, University of Minho.