Journal of Speech Sciences 2(2):113-136.2012.
Available at: http://www.journalofspeechsciences.org

Extending Automatic Transcripts in a Unified Data Representation towards a Prosodic-based Metadata Annotation and Evaluation

BATISTA, F.[1,2]; MONIZ, H.[1,3]; TRANCOSO, I.[1,4]; MAMEDE, N.[1,4]; MATA, A. I.[3]

[1] L[2]F – INESC-ID, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal

[2] DCTI, ISCTE-IUL, Lisboa, Portugal

[3] FLUL/CLUL, University of Lisbon, Lisboa, Portugal

[4] IST - UTL, Lisboa, Portugal

Abstract

This paper describes a framework that extends automatic speech transcripts in order to accommodate relevant information coming from manual transcripts, the speech signal itself, and other resources, like lexica. The proposed framework automatically collects, relates, computes, and stores all relevant information together in a self-contained data source, making it possible to easily provide a wide range of interconnected information suitable for speech analysis, training, and evaluating a number of automatic speech processing tasks. The main goal of this framework is to integrate different linguistic and paralinguistic layers of knowledge for a more complete view of their representation and interactions in several domains and languages. The processing chain is composed of two main stages, where the first consists of integrating the relevant manual annotations in the speech recognition data, and the second consists of further enriching the previous output in order to accommodate prosodic information. The described framework has been used for the identification and analysis of structural metadata in automatic speech transcripts. Initially put to use for automatic detection of punctuation marks and for capitalization recovery from speech data, it has also been recently used for studying the characterization of disfluencies in speech. It was already applied to several domains of Portuguese corpora, and also to English and Spanish Broadcast News corpora.

Keywords: Automatic speech processing, speech alignment, structural metadata, speech prosody, speech data representation, multiple-domain speech corpora, cross-language speech processing.

1 Introduction

Automatic speech recognition systems (ASR) are now being applied to a vast number of speech sources, such as radio or TV broadcasts, interviews, e-learning classes. Improving the output of a speech recognition system involves, not only reducing the well-known word error rate (WER), but also involves the process of recovering structural information and the creation of metadata from that information in order to produce richer transcripts. Enriching speech transcripts with structural metadata levels is of crucial importance and comprises several metadata extraction/annotation tasks,

* Corresponding authors: {fmmb,helenam,imt,njm}@l2f.inesc-id.pt, aim@fl.ul.pt

such as: speaker diarization, which consists of assigning the different parts of the speech to the corresponding speakers; sentence segmentation or sentence boundary detection; punctuation recovery; capitalization recovery; and disfluency detection and filtering. Such metadata extraction/annotation technologies are recently receiving increasing importance (1-3), and demand multi-layered linguistic information to perform such tasks.

Manually transcribed speech provides reference data that can be useful for speech analysis, for training, or for evaluating certain speech processing tasks. 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, capitalization, metadata indicating the presence of foreign languages, and other phenomena, such as disfluency marking. Automatic transcripts, produced by automatic speech recognition (ASR) systems, differ from manual transcripts in many ways: this type of data usually corresponds to a sequence of lowercase words, each of which referring its corresponding time period in the speech signal 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. Apart from the previously mentioned data, prosodic information (pitch and energy) is usually not available in the speech transcripts, but can be directly extracted from the speech signal. Combining the automatically produced speech information with additional elements, which may be manually provided or automatically calculated from other sources, is essential for a number of speech tasks that deal with ASR data. Moreover, different ASR systems produce different transcripts and also, because ASR systems are constantly being improved, different versions of the same ASR system will certainly produce transcripts that diverge in many aspects. One way of dealing with one of these outputs for a given task is to build a task dependent complex program that performs the link between the ASR output and the other available information sources. Another way to successfully deal with these different outputs is to automatically transfer the existing manual annotations, as well as other available elements, into the automatic transcript itself, in a way that small programs can easily process it. A self-contained dataset that merges all the information and can be easily dealt with constitutes a valuable resource for extensively addressing, studying and processing speech data.

Recovering punctuation marks, capitalization and disfluencies are three relevant MDA (Metadata Annotation) tasks. The impact of the methods and of the multi-layered linguistic information on structural metadata tasks has been discussed in the literature. (4, 5) report a general HMM (Hidden Markov Model) framework that allows the combination of lexical and prosodic clues for recovering full stop, comma and question marks. A similar approach was also used by (2, 6, 7) for detecting

sentence boundaries. (5) also combines 4-gram language models with a CART (Classification and Regression Tree) and concludes that prosodic information highly improves punctuation generation results. A Maximum Entropy (ME) based method is described by (8) for inserting punctuation marks into spontaneous conversational speech, where the punctuation task is considered as a tagging task and words are tagged with the appropriate punctuation. It covers three punctuation marks: comma, full stop, and question mark; and the best results on the ASR output are achieved by combining lexical and prosodic features. A multi-pass linear fold algorithm for sentence boundary detection in spontaneous speech is proposed by (9), which uses prosodic features, focusing on the relation between sentence boundaries and break indices and duration, covering their local and global structural properties. Other recent studies have shown that the best performance for the punctuation task is achieved when prosodic, morphologic and syntactic information is combined (2, 3, 10-12).

Much of the features and the methods used for sentence-like unit detection are applied in disfluency detection tasks. What is specific of the latter is that disfluencies obey to a structure: reparandum interruption point, interregnum and repair of fluency (13-15). 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 it may have 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 (2, 13, 14, 16-18). There is in fact an *edit signal* process (18), meaning that speakers signal an upcoming repair to their listeners. The edit signal is manifested by means of repetition patterns, production of fragments, glottalizations, co-articulatory gestures and voice quality attributes, such as jitter (perturbations in the pitch period) in the reparanda. Sequentially, it is also edited by means of significantly different pause durations from fluent boundaries and by specific lexical items in the interregnum. Finally, it is edited via f_0 and energy increases in the repair. The main focus is thus to detect the interruption point or the frontier between disfluent and fluent speech. For the interruption point detection task (2) reports that prosody alone produces best results on downsampled data, and combining all features produces best results on non-downsampled.

This paper proposes a framework that extends existing ASR transcripts in order to accommodate other relevant information concerning the speech data, coming from several sources, including the correspondent manual transcripts, the speech signal itself, and other speech related resources, like lexicons. The proposed framework collects, relates, computes, and stores all relevant information together, in a single data source, making it easier to deal with a wide range of interconnected information. The framework involves an automatic processing chain, where the first construction stage consists of integrating relevant reference data, coming from manual annotations, into the ASR output. The second and final stage consists of providing additional lexical, acoustic and prosodic related

information into the existing data. Aligning manual and automatic transcripts is not a trivial task mainly because of the recognition errors, but also due to the presence of complementary information. The paper gives a special focus to these alignment issues and reports alignment experiments of three metadata elements: punctuation, capitalization and disfluencies, in a base version (section 3) and in an improved version using prosodic features (section 5).

The resulting content has been extensively used for speech analysis, and also for training and testing semi-supervised learning methods for automatic punctuation and capitalization recovery over automatic speech data. The first studies were performed on Portuguese Broadcast News, but more recent studies also cover English and Spanish Broadcast News, and two different domains of Portuguese data: university lectures and map-task corpora. Recent steps have been given to use the unified data representation for the creation of ASR models for elderly, on the scope of the Project AVoz (FCT/PTDC/EEA-PLP/121111/2010), and also for the analysis of prepared non-scripted and spontaneous speech in school context, on the scope of the Project COPAS (FCT/PTDC/CLE-LIN/120017/2010).

This paper is organised as follows: Section 2 describes the scope of already performed experiments and some of the most relevant corpora used in this paper. Section 3 reports the first level of integration between the manual annotations and the speech recognition output. Section 4 describes the process of providing prosodic information to the data. Section 5 describes some variants and applications of the proposed framework. Finally, in Section 6, conclusions are presented.

2 Current scope and data

Aiming at a full account of all the levels of structural metadata in the ASR and the ways in which they interact, work has been conducted to enrich speech transcripts with capitalization, punctuation marks and disfluencies, in line with other work reported in the literature (2, 3). This framework has been developed and improved with the initial purpose of being the starting point for experiments with automatic punctuation and capitalization recovery, but most recent studies have also taken advantage of this framework for performing a vast number of speech studies on disfluencies. The first studies were performed on Portuguese and English Broadcast News (BN), but more recently two additional domains of Portuguese data were also studied: university lectures (19) and map-task corpora (20). While BN was used mostly for the punctuation and capitalization tasks, the university lectures, richer in spontaneous speech, were preferred for studying and analysing disfluencies. The following paragraphs describe the corpora that had been used in our experiments.

The ALERT-SR speech corpus is an European Portuguese Broadcast News corpus, originally collected from the public TV channel (RTP) for training and testing the speech recognition and topic detection systems, in the scope of the ALERT European project (21, 22). The training data

corresponds to about 47h of useful speech and contains about 480k words. The evaluation set was complemented with two collections of 11 BN shows, now totalling 26h of data. The manual orthographic transcription process follows the LDC Hub4 (Broadcast Speech) transcription conventions, and includes information such as punctuation marks, capital letters and special marks for proper nouns, and acronyms. Each segment in the corpus is marked as: planned speech with or without noise (F40/F0); spontaneous speech with or without noise (F41/F1); telephone speech (F2); speech mixed with music (F3); non-native speaker (F5); any other speech (FX). Most of the corpus consists of planned speech, but it also contains a considerable percentage (35%) of spontaneous speech.

The LECTRA corpus was collected within the homonym national project that aimed at transcribing lectures for the production of multimedia lecture contents for e-learning applications, and also for enabling hearing-impaired students to have access to recorded lectures (19). The corpus includes six courses in Portuguese and recorded in the presence of students and one course that was recorded in a quiet environment, targeting an Internet audience. The initial 21h orthographically transcribed were recently extended to 32h (23), under the scope of the European Project METANET4U. 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 in both development and test sets. Our experiments use the training portion of the corpus.

The English BN corpus used in our experiments combine five different English BN corpora subsets, available from the Linguistic Data Consortium (LDC). From the corpus LDC1998T28 (HUB4 1997 BN training data), about 94% was used for training and the rest for development. The first 80% of the LDC2005T24 corpus (RT-04 MDE Training Data Text/Annotations) was used for training, 10% for development and the last 10% for evaluation. The evaluation data also includes the LDC corpus LDC2000S86 (HUB4 1998 BN evaluation), LDC2000S88 (HUB4 1999 BN evaluation), LDC2005T24 (MDE RT04, only the last 10% were used), and LDC2007S10 (NIST RT03 evaluation data). The final corpus contains 81h (transcribed speech only) of training data, 6h of development data, and 9h of data for evaluation. The training data is almost twice the size of the Portuguese BN training data. Dealing with such corpora demanded normalization strategies, specifically adapted for each corpus. They have been produced in different time periods, encoded with different annotation criteria, and are available in different formats as well. Besides, they were built for different purposes, which make them even more heterogeneous. One important step for dealing with these corpora consisted on converting the original format of each one into a common format. The chosen common format was the STM (Segment Time Mark) format, which is easy to process and understand, and can easily map all the information required for our experiments. These corpora contain portions of overlapped speech but, in order to correctly use our recognition system, only one speaker was kept for such segments.

Besides the Portuguese and English data, a Spanish corpus (24) has also been used in some of our experiments (25). It consists of 20h of manually annotated news from the national Spanish TV station (TVE), collected between 2008 and 2009.

For each corpus, we have both manual transcripts and transcripts produced by Audimus (26), our ASR system. The ASR system is able not only to produce automatic transcripts from the speech signal, but also to produce automatic force-aligned transcripts by means of adjusting the manual transcripts to the speech signal. Force-aligned transcripts depend on a manual annotation and therefore do not contain recognition errors. On the other hand, automatic transcripts usually include ASR errors, highly dependent of the corpora type, speaker fluency, acoustic conditions, etc. Force-aligned transcripts are sometimes preferable to fully automatic transcripts for model training, due to the absence of speech recognition errors. For certain tasks, such as capitalisation that do not require information only available in the ASR output, the manual transcripts could be used directly. However, a number of speech tasks, such as the punctuation recovery task, use important information, such as pause durations, which most of the times are not available in the manual transcripts. For that reason, force--aligned transcripts must be used instead of manual transcripts. An important advantage of using force-aligned transcripts is that they can be treated in the exact same way as the automatic transcripts, but without recognition errors, requiring the same exact procedures and tools.

3 Integrating reference data in the ASR output

The proposed framework aims at producing self-contained datasets that can provide not only the information given by the ASR system, but also all the required reference data and other relevant information that can be computed from the speech signal. The original information provided by our ASR system is represented using standard XML (Extensible Markup Language). The resulting information is also stored in the standard XML representation format, thus simply extending the original DTD (Document Type Definition) file. One of the reasons for adopting the XML format is that all the already in use speech processing tools and modules (*e.g.*, topic detection, speech summarization, and multimedia content production modules) can still use the extended version without modifications, and can easily be upgraded to account for the additional information. Another reason is that processing XML data is now extremely facilitated by the existing efficient XML APIs (Application Programming Interfaces) that can be found for almost any programming language. It is now possible to process arbitrarily complex data structures, represented in XML format, with less than a few programming lines. Other open source representation formats could be explored. For example, NiteXML (27) is a powerful representation format, also based on XML, which comes with end user graphical interfaces for performing common tasks, and is powerful enough for representing different annotation layers and for relating them with the speech signal. For example, NiteXML was used to create the NXT-format Switchboard Corpus, a resource that combines many annotations of the

Switchboard corpus (28). However, rather than considering alternate representation formats, our focus is in automatically combining information from different sources into a self-contained data set, with the purpose of being easily addressed by small programs. That allows easily producing large amounts of statistical information that can be analysed and used for a large number of tasks.

As previously mentioned, this work was initially developed in order to produce suitable data for training and evaluating two speech processing tasks: automatic recovery of punctuation marks, and automatic capitalisation. Manual orthographic transcripts, usually performed at the segment level, constitute our reference data, and include punctuation, capitalization, and disfluency information. That is not the case of the fully automatic and force-aligned transcripts, which includes time intervals for other units of analysis (different segments, words, and possibly phones or syllables) together with confidence scores (of extreme importance for model training) but does not contain the reference data. The required reference must then be provided to the ASR output by means of alignments between the manual and automatic transcripts, a non-trivial task due to recognition errors. The initial stage in the proposed framework consists of transferring all relevant manual annotations to the automatically produced transcripts. The remainder of this Section describes the pre-processing steps and reports statistics about the alignment issues concerning the previously mentioned tasks.

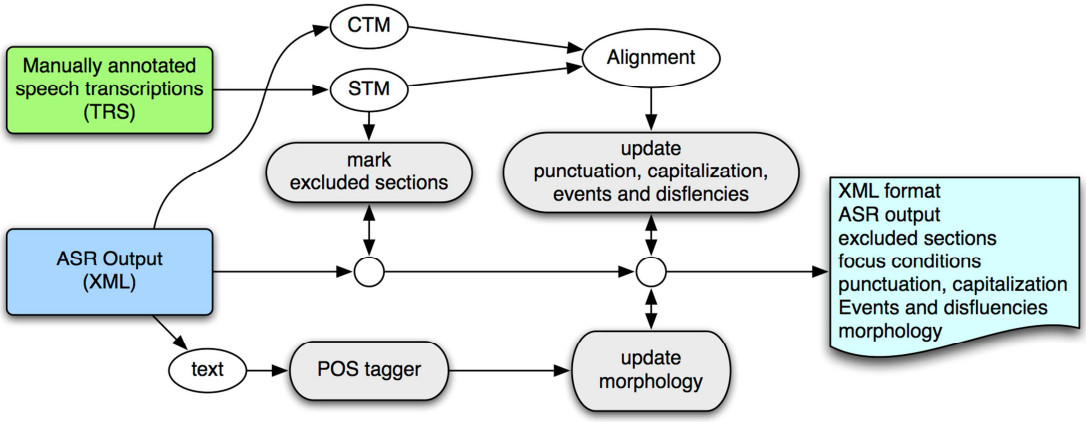

Figure 1 - Creating a file that integrates the reference data into the ASR output.

Figure 1 illustrates the process of integrating the reference data in the automatic transcripts, and providing additional meta-information to the data. The alignment process requires conversion of the manual transcripts, usually available in the TRS (XML-based standard Transcriber) format, to the STM (segment time mark) format, and the automatic transcripts into CTM (time marked conversation scoring). The STM format assigns time information at the segment level, while the CTM format assigns it at the word level. The alignment is performed using the NIST SCLite tool (http://www.nist.gov/speech) followed by an automatic post-processing stage, for correcting possible SCLite errors, and aligning special words, which can be written/recognized differently and are not considered by SCLite. The automatic post-processing stage, performed after the SCLite alignment,

allows overcoming problems, such as words A.B.C. or C.N.N. appearing as single words in the reference data, but recognized as isolated letters.

All corpora are also automatically annotated with part-of-speech information. The part-of-speech tagger input corresponds to the text extracted from the ASR transcript, after being improved with the reference capitalization. Hence, a generic part-of-speech tagger that processes written texts can be also used to perform this task, taking into account the surrounding words. However, it is important to stress that the performance will be decreased due to recognition errors. Currently, the Portuguese data is being annotated using *Falaposta*, a CRF-based tagger robust to certain recognition errors, given that a recognition error may not affect all its input features. It performs a pre-detection of numbers, roman numerals, hyphens, enclitic pronouns, and foreign words, and uses features based on prefix, suffix, case info, and lowercase trigrams. It accounts for 29 part-of-speech tags, processes 14k words/second, and it achieves 95.6% accuracy. The Spanish and the English data are being annotated using *TreeTagger* (29), a language independent part-of-speech tagger, which was used with the included standard dictionaries.

```
<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 2 - Example of an ASR transcript segment, enriched with reference data.

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 part-of-speech information. Figure 2 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. As for the attributes of the segment, it is described with the following sequential information: it is a segment identified as

12, it has been automatically characterized as "clean" by the ASR system with a confidence level (0.548); the segment temporal interval has been delimited with a very high confidence level (0.995); the speaker identity is "2001" with a low confidence in the recognition of this specific identity (0.379), but a high one in the classification of its gender as a female (0.897); and it is a native speaker of Portuguese. As for the wordlist itself: 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 (*e.g.*, "A" for the adjective *"Boa"/Good*, "Np" for proper noun "Mourinho").

The punctuation, part-of-speech, focus conditions, and information concerning excluded sections were updated with information coming from the manual transcripts. However, the reference transcript segments, manually created and more prone to be based on syntactic and semantic criteria, are usually different from automatically created segments, given by the ASR and APP (Audio pre-processing) modules, which are purely based on the signal acoustic properties. Therefore, whereas, in the reference, exclusion and focus information are properties of a segment, in the ASR output such information must be assigned to each word individually.

Recent efforts have also focused on including paralinguistic information (breath pausing, laughs, etc.) and disfluencies, both contained in the manual references, into the resultant enriched ASR file. Figure 2 shows examples of instantaneous events, such as jingles (Event name="[JINGLE_F]") and inspirations (Event name="[I]") that appear between the words. However, other metadata annotation, covering phenomena like disfluencies, is also being incorporated in the final output. Figure 3 illustrates an example of disfluency, containing the filled pause <%aa>. The sentence corresponds to: "*estávamos a falar %aa de espaços lineares*"/ *we were talking about uh linear spaces*. While marking instantaneous events corresponds to inserting a single event element, disfluency marking deserves special attention because it delimits regions, which may or may not involve more than one segment. Incorporating events, like disfluencies, make it possible to have other levels of analysis in a constant enrichment of the transcripts.

```
<TranscriptWordList>
 <Word start="3829" end="3865" conf="0.997228" focus="F1" pos="V.">estávamos</Word>
 <Word start="3866" end="3869" conf="0.997015" focus="F1" pos="S.">a</Word>
 <Word start="3870" end="3919" conf="0.999482" focus="F1" pos="V.">falar</Word>
 <Event name="BEGIN_disf"/>
 <Word start="3943" end="3975" conf="0.685290" status="filled_pause" focus="F1">%aa</Word>
 <Event name="END_disf"/>
 <Word start="3976" end="3989" conf="0.997181" focus="F1" pos="S.">de</Word>
 <Word start="3990" end="4039" conf="0.282515" focus="F1" pos="Nc">bases</Word>
 <Word start="4213" end="4230" conf="0.998758" focus="F1" pos="S.">de</Word>
 <Word start="4231" end="4269" conf="0.999629" focus="F1" pos="Nc">espaços</Word>
 <Word start="4270" end="4316" conf="0.999783" focus="F1" punct="." pos="A.">lineares</Word>
</TranscriptWordList>
```

Figure 3 – Excerpt of an enriched ASR output with marked disfluencies.

As previously mentioned, integrating the reference data into the ASR output is performed by means of word alignments between the two types of transcripts, which is a non-trivial task mainly because of the recognition errors. The following two subsections describe the options that have been taken, and report the level of success accomplished in performing this integration.

3.1 Capitalization alignment

The main issue about providing a reference capitalization to the ASR output is that, for each and every word in the ASR output, a capitalization form must be assigned. That means that if a mistake is made, the resultant data will provide a poorer reference, which will be reflected in further speech processing. When in the presence of a correct word, the capitalization can be assigned directly. However, problems arise when in the presence of recognition errors.

```
1)
   REF: noutro processo também em Portugal que      está junto que é um apenso dos autos
   HYP: noutro processo também ** ******** portugal está junto que é um apenso nos alpes
2)
   REF: O pavilhão desportivo do Colégio Dom     Nuno Álvares Pereira
   HYP: o pavilhão desportivo do ******* colégio dono novas   pereira
3)
   REF: A SAD a      administração da SAD Luís Duque e     Augusto Inácio
   HYP: * lhe assada administração da *** sad  luís  duque augusto inácio
4)
   REF: Esta noite em Gondomar o líder dos Social Democratas
   HYP: esta noite em gondomar o líder dos ****** social-democratas
```

Figure 4 - Capitalization alignment examples.

Figure 4 shows examples of word alignments, extracted from the SCLite output, where misalignments are marked using colours. The first line of each example corresponds to the manual annotations (REF), while the second line (HYP) contains only lowercase words and corresponds to the ASR output. Both lines are forced to contain the exact same number of tokens, for that reason asterisks are used to insert empty words. Given the alignments, the goal is to transfer the capitalisation information from REF to HYP. The capitalization of some of these words is trivially transferred (*e.g.*, Pereira, Augusto, Gondomar). The capitalisation of other words, *e.g.*, "Portugal" and "Colégio"/*College*, can also directly be solved by the automatic post-processing stage, by looking at the words in the neighbourhood. The capitalization of compounded words, like "social-democratas"/*social-democrats* in the example 4, is also being treated in the post-processing stage. In fact, all the underlined words become capitalized after applying the post-processing step. If no information exists concerning the capitalization of a word, it is considered lowercase by default. Therefore, any word inserted by the recognition system that does not exist in the reference (insertion) will be kept lowercase. Similarly, if a reference word was skipped by the recognition system (deletion), nothing can be done about it. Anyway, most of the insertions and deletions consist of short functional words, which usually appear in lowercase and do not pose significant problems to the reference capitalization. Finally, if the words mismatch but the reference word is lowercase, the word in the automatic transcript is also marked as lowercase and will not pose problems to the reference capitalization either.

Most of the alignment problems arise from word substitution errors, where the reference word appears capitalized (not lowercase). In this case, three different situations may occur: i) the two words have alternative graphical forms, a not infrequent phenomena in proper nouns, for example: ”Menezes” and “Meneses” (proper nouns); ii) the two words are different but share the same capitalization, for example: “Andreia” and “André”; and iii) the two words have different capitalization forms, for example “Silva” (proper noun) and “de” (of, from). As the process is fully automatic, we have decided to assign the same capitalization information whenever the Levenshtein distance (30) between the two words was less than 2. By doing this, capitalization assignments like the following were performed: “Trabalhos” → “Trabalho”; “Espanyol” → “Espanhol”; “Millennium” → “Millenium”; “Claudia” → “Cláudia”; “Exigimos” → “Exigidos”; “Andámos” → “Andamos”; “Carvalhas” → “Carvalho”; “PSV” → “PSD”; “Tina” → “Athina”. Notice that if the capitalization assignments of the above words were not performed, those words would appear lowercase in the reference capitalization, which would not be correct.

Table 1 - Capitalization alignment report.

	Cor	Del	Ins	lower case subs	Corrected alignments			Unsolved alignments		WER
					Sclite probs	Compound words	subs	First cap	All upper	
Train	87%	1.9%	4.5%	5.5%	0.5%	0.1%	0.3%	0.6%	0.0%	13.6%
Development	81%	2.5%	5.5%	8.4%	0.7%	0.1%	0.5%	0.9%	0.1%	19.0%
Eval	81%	2.6%	5.5%	8.5%	0.5%	0.0%	0.4%	1.1%	0.1%	19.4%
Jeval	82%	3.4%	4.4%	7.8%	0.6%	0.1%	0.5%	1.1%	0.1%	18.1%
Rtp07	81%	2.2%	6.3%	8.4%	0.6%	0.1%	0.4%	1.0%	0.1%	19.9%
Rtp08	76%	2.3%	10.3%	9.7%	0.5%	0.0%	0.5%	1.1%	0.1%	26.8%

Table 1 presents statistics concerning the capitalization assignment after the word alignment, for different parts of the ALERT-SR corpus. The proportion of correct alignments is shown in column *Cor*; *Del* and *Ins* correspond to the number of deletions and insertions in the word alignment; *lowercase subs* correspond to substitution of words involving lowercase words, which do not pose problems to correctly transfer the capitalization information. *Corrected alignments* show the percentage of corrections performed during the post-processing stage. The *unsolved alignments* correspond to unsuccessful alignments, involving first capitalized words (*e.g.*, proper nouns), and all uppercase letters (*e.g.*, acronyms). Statistics concerning other types of capitalization (*e.g.*, McDonald's) are not reported, because these words are rarely found, especially in speech data. The recognition Word Error Rate (WER) for each subset is shown in the last column. The proportion of recognition errors in each corpus subset is highly correlated with the other presented values. For example, higher WER values correspond to lower *Cor* values and higher *unsolved alignments* values.

3.2 Punctuation Alignment

Like the reference capitalization, inserting the correct reference punctuation into the automatic transcripts is not an easy task, though it poses different challenges.

```
1)
  REF: ESTAMOS SEMPRE A     DIZER À      senhoria .
  HYP: ******* ****** CALMO SEM   PESARÁ senhoria *
2)
  REF: no centro , O rio ceira ENCHEU de forma que A aldeia de CABOUCO ficou INUNDADA .
  HYP: no centro * * rio ceira INÍCIO de forma que * aldeia de TEMPO   ficou ******** *
3)
  REF: é a primeira vez que isto LHE acontece ?
  HYP: é a primeira vez que isto *** acontece *
4)
  REF: sem PERCEBEREM , SEM       LHES  DIZEREM quais são as consequências desta política
  HYP: sem ********** * RECEBEREM SELHO DIZER   quais são as consequências desta política
5)
  REF: ALIÁS ,  alguém DISSE , E EU ESTOU    de acordo , que hoje não temos UM governo ,
  HYP: HÁLIA ÀS alguém ***** * * ** INDÍCIOS de acordo * que hoje não temos O  governo *
6)
  REF: no segundo * **** ,   COLIN MONTGOMERY ,     JARMO SANDELIN , michael e laura
  HYP: no segundo O QUAL NÃO COBRE E          CRIAR UMA   CÉLULA   E michael e laura
```

Figure 5 - Punctuation alignment examples.

Figure 5 shows word alignment examples extracted from the SCLite output, where the matches are represented in blue. The first line of each example corresponds to the manual annotations (REF), while the second line (HYP) contains the ASR output. The goal consists of transferring all the punctuation information from REF to HYP. It is important to notice that the effect of the speech recognition errors is only relevant when they occur in the neighbourhood of a punctuation mark. Also notice that transferring all the punctuation marks is fundamental, despite the ASR errors, because a punctuation mark is not only related with the word itself but also with acoustic and prosodic information found in the speech signal. Not transferring a given punctuation implies associating features that usually characterize a punctuation mark to non-punctuation. The recognition errors in the first three examples do not pose problems to transferring the reference punctuation, because each punctuation mark is in the neighbourhood of a matching word. For example, in the second alignment, the *comma* follows the word "centro"/*center* and the *full-stop* follows the word "ficou"/*was*, despite of the missing words. However, the last three examples present more difficult challenges. The fourth and the fifth examples can still be solved in an acceptable manner, and provide acceptable reference data, even though it becomes very difficult to read. For example, the outcome for 4) would be: "sem, receberem selho dizer quais são as consequências desta política"/*without, receiving* selho *to say what are the consequences of this policy*, and the outcome for 5) would be "Hália às alguém, indícios de acordo, que hoje não temos o governo,"/ *Halia at someone, evidence of an agreement, that today we have no government,*. The last example is very difficult to solve and will probably provide a bad reference data, because the manual transcripts data indicate the use of three commas, but the ASR output mismatches the reference for a sequence of words. Notice that even a human annotator would consider difficult to transfer the three commas into the HYP sentence.

Table 2 - Punctuation alignment report.

Corpus subset	full-stop (.)			comma (,)			question mark (?)		
	Good	Ok	Bad	Good	Ok	Bad	Good	Ok	Bad
Train	71.4%	24.7%	3.9%	76.1%	18.4%	5.5%	41.1%	44.3%	14.7%
Development	66.3%	28.8%	4.9%	66.2%	27.1%	6.7%	33.7%	40.6%	25.7%
Eval	63.9%	30.1%	6.0%	64.6%	27.6%	7.9%	27.5%	51.0%	21.6%
Jeval	65.0%	30.7%	4.3%	65.8%	27.4%	6.8%	44.7%	37.5%	17.8%
Rtp07	63.6%	29.2%	7.2%	65.0%	25.6%	9.5%	27.5%	44.1%	28.4%
Rtp08	56.9%	30.9%	12.3%	59.8%	28.1%	12.1%	20.4%	49.0%	30.6%

Table 2 presents the alignment summary for the three most frequent punctuation marks, where the alignments are classified as *good*, *ok* (acceptable), or *bad*. The alignment performance is affected by the recognition errors, which are common in spontaneous speech. The worst values concern to the question mark, which is due to the fact that interrogative sentences appear more often in spontaneous speech. The final alignment would benefit from a manual correction, an issue to be addressed in the future. Nevertheless, as previously mentioned, even an expert human annotator would find difficulties in doing this task and sometimes would not perform it coherently.

3.3 Disfluencies and other events

Transferring disfluency marks and other events to the automatic transcript faces the same difficulties that have been described previously, which are caused mostly by the ASR errors. We have performed this task for BN data. However, the university lectures, richer in spontaneous speech, were preferred for studying and analysing disfluencies. As a consequence, instead of studying the fully automated speech recognition output, our experiments are based on force-aligned transcripts. That means that, in this case, instead of facing the challenges caused by the recognition errors, transferring the information from the manual to the automatic transcripts was relatively easy to perform, as the low alignment error clearly shows (0.9%; 1.3% and 1.6% for train, development and test sets, respectively).

```
1)
  REF: pode ser instanciado < no modelo de progra- > num modelo programado
  HYP: pode ser instanciado * no modelo de progra- * num modelo programado
2)
  REF: [BB] é interessante ou [BB] era uma dúvida [BB] pois < %aam > [TX] então
  HYP: **** é interessante ou **** era uma dúvida **** pois * %aam * **** então
3)
  REF: HUM ESTES ESTÃO AQUI CALADINHOS A ACEITAR O QUE ESTOU A DIZER
  HYP: *** ***** ***** **** ********** * ******* * *** ***** * *****
4)
  REF: TENHO ESTA CLASSE E A OUTRA E ESTAS DUAS INTERAGEM DESTA MANEIRA E < OU > A TERCEIRA
  HYP: ***** **** ****** * * ***** * ***** **** ********* ***** ******* * * ** * * ********
```

Figure 6 – Disfluency and other events alignment examples.

Figure 6 shows distinct examples of disfluency and other events (mis)alignment, extracted from the SCLite output. Angular brackets delimit disfluent sequences and other events are marked with square brackets, *e.g.*, [BB] (labial and coronal clicks) and [TX] (stands for cough). The complete list of events can be found in the *Transcriber* menus. Examples 1 and 2 show that, apart from the filled pause (%aam/*um*), the remaining events are not present in the force-aligned data. Examples 3 and 4 correspond to low energy segments that the ASR was unable to force-align. Example 3 is an aside from the teacher – *hum these here are so quiet accepting what I'm saying*. Whereas example 4

corresponds to an explanation of a slide with the teacher's head movement distorting the capture of the speech signal – *I have this class and other and these two interacting this way and <or> a third* [*class*].

4 Adding Prosodic Information

The generated XML, described in section 3, serves as a good data source for a number of experiments that rely purely on both lexical and audio segmentation features. However, the integration of the multilayer grammatical knowledge must also account for prosodic information. We have been saying that the motivation for the framework here described is mainly for structural metadata detection and analysis (2, 3). The view that prosody does play a role in sentence processing is detailed in (31), which gives us the following working definition of prosody:

"we specify prosody as both (1) acoustic patterns of F_0, duration, amplitude, spectral tilt, and segmental reduction, and their articulatory correlates, that can be best accounted for by reference to higher-level structures, and (2) the higher-level structures that best account for these patterns ... it is 'the organizational structure of speech'." (Shattuck-Hufnagel & Turk, p. 196)

In the above definition, prosody has two components: firstly, its acoustic correlates and, secondly, their relation to the organizational structure of speech. Detailed analyses have been conducted to describe the properties of the prosodic constituents and their functions, *e.g.* (32-41). Despite the different proposals and terminologies, there are cross language acoustic correlates to delimit sentence--like units (42). Features such as pause at the boundary, pitch declination over the sentence, postboundary pitch and energy resets, preboundary 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, as described in section 1.

This paper aims at providing a descriptive prosodic modelling for several domains and languages, thus it integrates the most informative acoustic correlates of sentence-like units detection for the specific metadata tasks (capitalization, punctuation and disfluencies), but it also provides a description of the levels of the prosodic structure for further applications. Thus, this section describes the prosodic feature extraction process and the creation of an improved data source, containing additional prosodic information, aiming at a view of the organizational structure of speech. In summary, Section 3 concerns all the alignment experiments (mostly the time alignment of words and events given by the ASR system); Section 4 describes the several steps taken to include prosodic features (beyond time alignment initially given by the ASR system); and Section 5 reports the positive impact of the prosodic features added to account for the specific metadata tasks.

4.1 Adding Phone Information

Audimus (26) 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 (MLP). Modelling context dependency is a particularly hard problem in hybrid systems. For that reason, this speech recognition system uses, in addition to monophone units modelled by a single state, 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 (43). The authors used a two-step method: first, a single state monophone model is extended to multiple state sub-phone modelling (*e.g.*, "L-b"; "b" and "b+R", where L stands for left state units and R for right state units); and secondly, a reduced set of diphone recognition units (*e.g.*, "d=i") is incorporated to model phone transitions. This approach is supported on the view that each phone is usually considered to be constituted by three regions or portions: an initial transitional region ("L-b"), a second central steady region, known as phone nucleus ("b"), and a final transitional region ("b+R"). The authors initially expected that modelling each one of these portions independently would improve the acoustic phone modelling. Their expectations were confirmed, leading to a reduction of 3% in the word error rate (from 26.8% to 23.8%). Figure 7 presents an excerpt of a PCTM input file, produced by the speech recognition system, and containing a sequence of phones/diphones, corresponding to the sequence: *"muito bom dia"/good morning*. The phonetic transcription uses SAMPA (Speech Assessment Methods Phonetic Alphabet).

```
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.00 0.27 interword-pause
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.27 0.01 L-m
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.28 0.01 m
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.29 0.04 m=u~
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.33 0.01 u~
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.34 0.02 u~=j~
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.36 0.01 j~
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.37 0.03 j~=t
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.40 0.01 t
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.41 0.02 t=u
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.43 0.01 u
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.44 0.01 u+R+
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.45 0.01 L-b
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.46 0.02 b
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.48 0.01 b+R
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.49 0.02 L-o~
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.51 0.05 o~
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.56 0.05 o~+R+
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.61 0.02 L-d
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.63 0.02 d
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.65 0.06 d=i
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.71 0.04 i
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.75 0.01 i=A
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.76 0.01 A
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.77 0.01 A+R+
2000_12_05-17_00_00-Noticias-7.spkr000 1 14.78 0.06 interword-pause
```

Figure 7 - PCTM file containing the phones/diphones produced by the ASR system.

The phones/diphones information was then converted into monophones by another tool, specially designed for that purpose. Such conversion process was accomplished due to an analysis performed in a reduced test set of 1h duration that was manually transcribed (44). The analysis of this sample revealed several problems, namely in the boundaries of silent pauses, and in their frequent misdetection, problems that affected the phone boundaries. Figure 8 presents an excerpt of the

resulting information. Still, the existing information is insufficient for correctly assigning phone boundaries. We have used the mid point of the phone transition, but setting more reliable monophone boundaries would, then, enable us to process pitch adjustments and, thus, to mark syllable boundaries and stress in a more sustainable way. For example, the phone sequence "j j=u u u+R", presented in Figure **9**, must be converted into the monophone sequence "j u", but the exact boundary between the first and the second phone is unknown.

```
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 8 - PCTM of monophones, marked with syllable boundary (the diacritic #) and stress (the diacritic ").

The speech recognition system could alternatively perform the speech recognition based purely on monophones, but then the cost would be reflected in an increased WER.

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. This was a problem, since we know from the extensive literature on prosodic studies 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. For the Portuguese BN, 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. Regarding the English language, most of the lexicon content was created from the CMU dictionary (version 0.7a). The phone sequence for the unknown words was provided by the text-to-phone CMU/NIST tool `addttp4`, and the stressed syllables were marked using the tool `tsylb2` (45), which uses an automatic phonological-based syllabication algorithm. The excerpt presented in Figure 8 shows an example of marked syllable boundaries and stress.

4.3 Extracting Pitch and Energy

Pitch (f_0) and energy (E) are two important sources of prosodic information that can be extracted directly from the speech signal. By the time these experiments were conducted, that information was not available in the ASR output. For that reason, it has been directly extracted from the speech signal, using the Snack (46) toolkit. Both pitch and energy were extracted using the standard parameters, taken from the `Wavesurfer` tool configuration (47). Energy was extracted using a pre-emphasis

factor of 0.97 and a hamming window of 200ms, while pitch was extracted using the ESPS method (auto-correlation). 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. Octave-jumps were also eliminated. As to the influences from noisy conditions, the recognizer has an Audio Pre-processing or Audio Segmentation module (22), which 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 9 illustrates the process described above, where the original pitch values are represented by dots and the grey line represents the resultant pitch. The first tier is the orthographic tier, also containing POS tags; the second tier corresponds to the multiple-state monophone/diphone units, and the last tier is the resulting conversion for monophones.

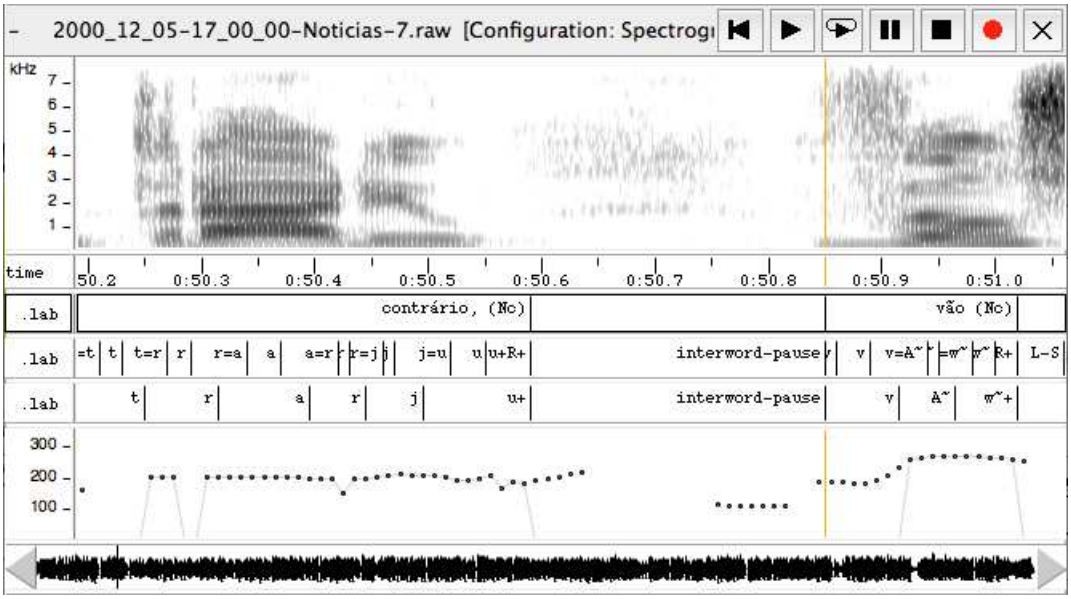

Figure 9 - Pitch adjustment.

4.4 Producing the final XML file

After extracting and calculating the above information, all data was merged into a single data source. The existing data source, previously described in Section 3, has been upgraded in order to accommodate the additional prosodic information.

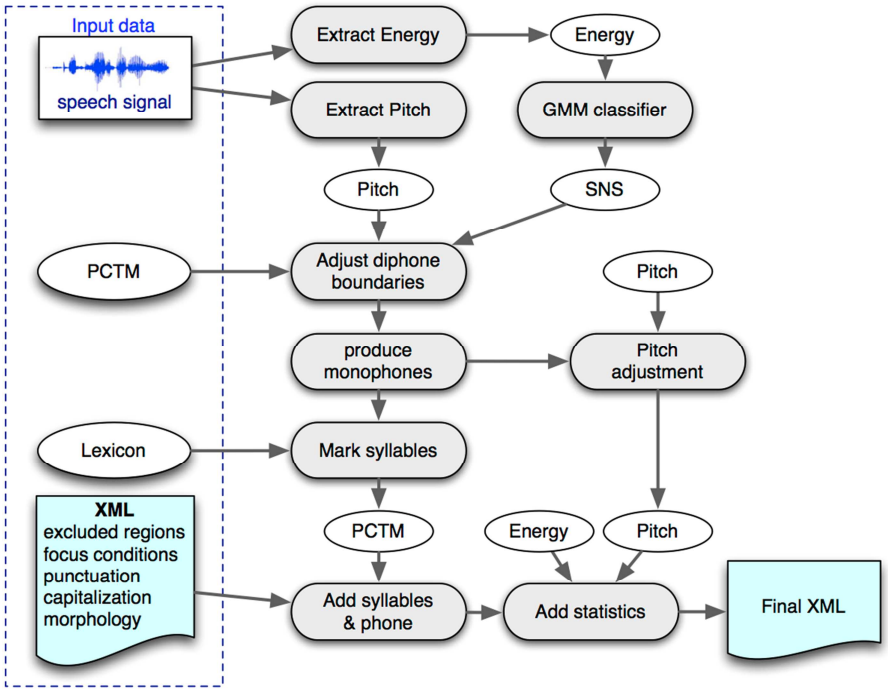

Figure 10 - Integrating prosodic information.

Figure 10 illustrates the involved processing steps for upgrading the existing data source with the additional information. The pitch and energy values are extracted directly from the speech signal. A Gaussian mixture model (GMM) 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. The PCTM file referred in Figure 7 is modified with the new unit boundaries, and then used to produce the file presented in Figure 8, containing only monophones. The monophone units are then used for removing the pitch values from unvoiced regions.

```
<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 11 – 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. In terms of DTD differences, apart from moving the word form from *text element* to an attribute of the element *Word*, the remaining changes consist of additional provided attributes. Figure 11 shows an excerpt from one of these files, 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. Pitch slopes were also calculated after converting the pitch into semitone values. Different measures are provided, relating to different base values: *pslope* (base is the minimum pitch in the range), *pmin_st_100* and *pmax_st_100* (base is 100Hz), *pmin_st_spkr* and *pmax_st_spkr* (base is the speaker minimum). Apart from that, the normalized values for pitch and energy have also been calculated: *pslope_norm* and *eslope_norm*.

5 Variants and applications

This section enumerates some applications that have taken advantage of the produced data and describes some variants of the proposed framework.

Two of those applications include automatic punctuation and capitalization recovery for automatic speech data, which are also part of the motivation for this work. The initial experiments concerning punctuation used only basic information, corresponding to the first stage of the proposed framework, described in Section 3. The most recent experiments involve a much more complex group of features that cover different units of analysis and make use of the prosodic properties of each one of these units (12). The paper reports different methods for improving the punctuation results. Different levels of linguistic structure, including lexical, prosodic, and speaker related features were used. Statistics were calculated for each word transition, with or without a pause, using: the last word, last stressed syllable and last voiced phone from the current word, and the first word, and first voiced phone from the following word. The following set of features has been used: f_0 and energy slopes in the words before and after a silent pause, f_0 and energy differences between these units and also duration of the last syllable and the last phone. This set of prosodic features proved useful for the detection of the full stop and comma on both languages, presenting a SER improvement from 3% to 8% (absolute), relative to the previous results obtained using only lexical and audio segmentation features. Thus, the best strategy, for both Portuguese and English, involved providing different levels of linguistic structure, including lexical, prosodic, and speaker related features. Despite the corpora being different, the presented results are similar to the ones reported by (2), concerning sentence boundary detection.

```
<TranscriptSegment start="3694" end="4316">
  <TranscriptWordList type="disfluency" start="3694" end="3753">
    </Word start="3694" end="3753" conf="0.996" focus="F1" pos="S." name="na">
  </TranscriptWordList>
  <TranscriptWordList type="chunk" start="3754" end="3919">
    </Word start="3754" end="3791" conf="0.999" focus="F1" pos="A." name="última">
    </Word start="3792" end="3824" conf="0.998" focus="F1" pos="Nc" name="aula">
    </Word start="3829" end="3865" conf="0.997" focus="F1" pos="V." name="estávamos">
    </Word start="3866" end="3869" conf="0.997" focus="F1" pos="S." name="a">
    </Word start="3870" end="3919" conf="0.999" focus="F1" pos="V." name="falar">
  </TranscriptWordList>
  <TranscriptWordList type="disfluency" start="3943" end="3975">
    </Word start="3944" end="3975" conf="0.685" status="filled_pause" focus="F1" name="%aa">
  </TranscriptWordList>
  <TranscriptWordList type="chunk" start="3976" end="4316">
    </Word start="3976" end="3989" conf="0.997" focus="F1" pos="S." name="de">
    </Word start="3990" end="4039" conf="0.282" focus="F1" pos="Nc" name="bases">
    </Word start="4213" end="4230" conf="0.998" focus="F1" pos="S." name="de">
    </Word start="4231" end="4269" conf="0.999" focus="F1" pos="Nc" name="espaços">
    </Word start="4270" end="4316" conf="0.999" focus="F1" punct="." pos="A." name="lineares">
  </TranscriptWordList>
</TranscriptSegment>
```

Figure 12 – Variant of the extended ASR output, using different segmentation units.

The proposed framework is also an important part of studies that have been recently performed on the prosodic properties of disfluencies and their contexts (48, 49). Such analysis required features calculated for the disfluent sequence and also for the two adjacent contiguous words. The adjustment to the new units is motivated by the fact that, when uttering a disfluency, there are distinct regions to account for (13-15), namely, the *reparandum*, or region to repair, the interruption point, the disfluency itself, and the repair of fluency. Due to these requirements, we have created a variant of the output produced after the framework's first stage, which simply reorganizes the information into different units. This variant uses the reference punctuation to assign each sentence to a *TranscriptSegment*, and considers two types of *TranscriptWordLists*. Figure 12 shows an example of a transcript segment containing four different groups of word lists, divided into "chunks" or "disfluency" types. The former comprises strings of words between disfluencies and the latter the discrimination of the disfluency itself or the sequence of disfluencies. By doing such modification, the second stage of our framework assigns the correct information to each one of these new units, allowing for the analysis of both local and global prosodic properties.

The XML file was applied as a substantial input to more thorough tempo, pitch and energy analysis of the disfluencies in a corpus of university lectures (48, 49). From all the layers of information detailed in the framework here described, the most informative features extracted to characterize the different regions at stake when producing disfluencies and when repairing it were: pitch and energy slopes, the differences between those slopes, and the tempo characteristics of the distinct regions and of the adjacent silent pauses. The statistical analysis of all the features allows reporting strong patterns in the data. Firstly, different regions of a disfluent sequence are uttered with distinct prosodic properties and speakers contrast, by means of prosodic marking, *i.e.*, pitch and energy increase, those areas with the minimum context possible (within the words immediately before and after a disfluency). Secondly, there are different contrastive degrees in using the prosodic parameters (filled pauses are the most distinct type in what regards pitch increase and durational aspects, and repetitions in what

concerns energy rising patterns). Finally, when repairing fluency, speakers overall produce both pitch and energy increases, but they monitor tempo aspects in an idiosyncratic way.

It is also important to mention that the unified ASR output representation considers and stores a minimal set of information, from which an extended set of features can be derived. For example, the following set of features were considered for the previously described task on analysis of disfluencies: a) number of words, syllables, and phones inside and outside disfluencies; b) duration of speech with and without utterance internal silences; a) articulation rate, rate of speech, and phonation ratio, both per speaker and per sentence, with and without disfluencies. Apart from these two major usages of the data produced by our framework, several other applications have indirectly used the extended ASR output produced by the framework, including machine translation and summarization tasks.

6 Conclusion

Producing rich transcripts involves the process of recovering structural information and the creation of metadata from that information. Enriching speech transcripts with structural metadata levels is of crucial importance and demands multi-layered linguistic information to accomplish those tasks. This paper describes a framework that permits to extend the output of a speech recognition system in order to accommodate for additional information coming from manual transcripts, the speech signal, or other information sources. Such framework was motivated by three relevant metadata annotation tasks: recovering punctuation marks, capitalization and disfluencies.

The process of transferring manual transcripts information into the ASR output constitutes the first stage in the described framework. That is accomplished by performing alignments between the manual and the automatic transcripts, a non trivial task due to recognition errors. The paper reports a number of alignment experiments concerning punctuation, capitalization, and paralinguistic information. The second stage in our framework consists of upgrading the previous integrated information, by introducing phone information, syllable boundaries, syllable stress, and other prosodic speech properties extracted from the signal. Each unit of analysis, including sentences, words, syllables, and phones became fully characterized with pitch and energy measures. During the process, pitch, energy and duration were used to adjust word boundaries, automatically identified by the speech recognition system, contributing to an improved characterization of the speech data.

The final content extends the original ASR output and is available as a self-contained XML file that can provide, not only the information given by the ASR system, but also all the required reference data and other relevant information computed from the speech signal. Such self-contained file can still be used with existing speech processing tools and modules without adaptation and can easily be processed by small programs in order to produce a wide range of statistical data.

7 Acknowledgments

Helena Moniz is supported by FCT grant SFRH/BD/44671/2008. This work was partially funded by CMU-PT/HuMach/0039/2008, by PTDC/CLE-LIN/120017/2010 COPAS, by PEst-OE/EEI/ LA0021/2011, and by DCTI - ISCTE-IUL – Lisbon University Institute.

8 References

1. Jurafsky, D. and Martin, J. H. (2009). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR, second edition.
2. Liu, Y., Shriberg, E., Stolcke, A., Dustin, H., 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, vol. 14, n. 5, pp. 1526-1540.
3. 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. and Wooters, C., (2008) Speech Segmentation and Spoken Document Processing. IEEE Signal Processing Magazine, pp. 59-69.
4. Christensen, H., Gotoh, Y., and Renals, S. (2001). Punctuation annotation using statistical prosody models. In Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding, pages 35–40.
5. 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.
6. Gotoh, Y. and Renals, S. (2000). Sentence boundary detection in broadcast speech transcripts. In Proc. of the ISCA Workshop: Automatic Speech Recognition: Challenges for the new Millennium ASR-2000, pages 228–235.
7. Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. Speech Communications, 32(1-2):127–154.
8. Huang, J. and Zweig, G. (2002). Maximum entropy model for punctuation annotation from speech. In Proc. of the 7th International Conference on Spoken Language Processing (INTERSPEECH 2002), pages 917 – 920.
9. Wang, D. and Narayanan, S. S. (2004). A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’04), volume 1, pages 525–528
10. Shriberg, E., Favre, B., Fung, J., Hakkani-Tur, D., and Cuendet, S. (2009). Prosodic similarities of dialog act boundaries across speaking styles. Linguistic Patterns in Spontaneous Speech - Language and Linguistics Monograph Series, 25:213–239.
11. Favre, B., Hakkani-Tur, D., and Shriberg, E. (2009). Syntactically-informed Models for Comma Prediction. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’09), Taipei, Taiwan
12. Batista, F., Moniz, H., Trancoso, I., Mamede, N. J., (2012) Bilingual Experiments on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts, IEEE Transactions on Audio, Speech, and Language Processing, IEEE Signal Processing Society, vol. 20, n. 2, pages 474 - 485, doi: 10.1109/TASL.2011.2159594
13. Nakatani, C., Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. Journal of the Acoustical Society of America (JASA), (95):1603–1616.
14. Shriberg, E. (1994). Preliminaries to a Theory of Speech Disfluencies. PhD thesis, University of California.
15. Levelt, W. (1989). Speaking. MIT Press, Cambridge, Massachusetts.
16. Shriberg, E. (2001). To ”‘errrr”’ is human: ecology and acoustics of speech disfluencies. Journal of the International Phonetic Association, (31):153–169.
17. Levelt, W., Cutler, A. (1983) “Prosodic marking in speech repair,” Journal of Semantics, no. 2.

18. Hindle, D. (1983) Deterministic parsing of syntactic non-fluencies. In Proc. of the 21st annual meeting of the Association for Computational Linguistics (A CL-83), pages 123-128.
19. Trancoso, I., Martins, R., Moniz, H., Mata, A. I., Viana, M. C. (2008). The LECTRA Corpus - Classroom Lecture Transcriptions in European Portuguese. In *Proc. LREC*, Marrakech
20. Trancoso, I., Viana, M. C., Duarte, I., Matos, G. (1998), Corpus de Diálogo CORAL, *In* PROPOR'98 - III Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, Porto Alegre, Brasil
21. Neto, J. P., Meinedo, H., Amaral, R., and Trancoso, I. (2003). A system for selective dissemination of multimedia information. In Proc. of the ISCA MSDR 2003
22. Meinedo, H. and Neto, J. P. (2003). Audio segmentation, classification and clustering in a broadcast news task. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’03), Hong Kong, China
23. Pellegrini, T., Moniz, H., Batista, F., Trancoso, I., Astudillo, R. (2012) Extension of the LECTRA corpus: classroom LECture TRAnscriptions in European Portuguese, In SPEECH AND CORPORA, Belo Horizonte
24. Meinedo, H, Abad, A., Pellegrini, T., Trancoso, I., Neto, J. P. (2010), The L2F Broadcast News Speech Recognition System, *In* Fala2010, Vigo, Spain
25. Batista, F., Trancoso, I., Mamede, N. J. (2009), Comparing Automatic Rich Transcription for Portuguese, Spanish and English Broadcast News, *In* ASRU - Automatic Speech Recognition and Understanding Workshop, Merano, Italy
26. Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., and Caseiro, D. (2008). Broadcast news subtitling system in Portuguese. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’08), pages 1561–1564
27. Carletta, J., Evert, S., Heid, U., and Kilgour, J. (2005). The NITE XML Toolkit: data model and query. Language Resources and Evaluation Journal 39(4): 313-334. DOI 10.1007/s10579-006-9001-9
28. Calhoun, S., Carletta, J., Brenier, J., Mayo, N., Jurafsky, D., Steedman, M. and Beaver, D. (2010) The NXT-format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue. Language Resources and Evaluation Journal 44(4): 387-419. DOI: 10.1007/s10579-010-9120-1
29. Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, United Kingdom.
30. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 6:707–710. (English translation).
31. Shattuck-Hufnagel, S. And Turk, A., A Prosody Tutorial for Investigators of Auditory Sentence Processing. Journal of Psycholinguistic Research, vol. 25, n. 2, pp. 193-247, 1996
32. Liberman, M. (1975). The intonational system of English. PhD Dissertation, MIT. Distributed 1978 by IULC.
33. Bruce, G. (1977). Swedish word accents in sentence perspective. Lund: Gleerup.
34. Pierrehumbert, J. (1980) The phonology and phonetics of English intonation. Ph.D. dissertation, MIT.
35. Beckman, M., Pierrehumbert, J. (1986). “Intonational structure in Japanese and English”. Phonology Yearbook, pp. 15-70.
36. Nespor, M., Vogel, I.(2007). Prosodic Phonology. Berlin/New York: Mouton de Gruyter. (2nd edition).
37. Gussenhoven, C. (2004). The Phonology of Tone and Intonation. Cambridge: Cambridge University Press.
38. Ladd, D. R. (1996). Intonational Phonology. Cambridge:CUP.
39. Ladd, D. R. (2008). Intonation Phonology, 2.ª Edição, Cambridge University Press, Cambridge.
40. Bolinger, D. (1989). Intonation and its uses: Melody in grammar and discourse. London:Arnold.
41. Pierrehumbert, J., Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In Philip R. Cohen, Jerry Morgan & Martha E. Pollack (eds.), Intentions in communication, 271-311. Cambridge, MA: MIT Press.
42. Vassière, J. (1983). Language-independent prosodic features. In Cutler, A. and Ladd, R., editors, Prosody: models and measurements, pages 55–66. Berlin: Springer

43. Abad, A. and Neto, J. (2008). Incorporating acoustical modelling of phone transitions in a hybrid ANN/HMM speech recognizer. In Proc. of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008), Brisbane, Australia
44. Moniz, H., Batista, F., Meinedo, H., Abad, A., Trancoso, I., Mata, A. I., Mamede, N. J. (2010) Prosodically-based automatic segmentation and punctuation, In Speech Prosody 2010, ISCA, Chicago, USA
45. Fisher, B. (1996). The tsylb2 program. National Institute of Standards and Technology Speech.
46. Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., and Granström, B. (1998). Web-based educational tools for speech technology. In Proc. of ICSLP98, 5th Intl Conference on Spoken Language Processing, pages 3217–3220, Sydney, Australia
47. Sjölander, K. and Beskow, J. (2000). Wavesurfer-an open source speech tool. In Sixth International Conference on Spoken Language Processing, pages 464–467
48. Moniz, H., Batista, F., Mata, A.I., and Trancoso, I., (2012) Analysis of disfluencies in a corpus of university lectures. In Proc. of Exling 2012, Athens, Greece.
49. Moniz, H., Batista, F., Trancoso, I., Mata, A. I., "Prosodic context-based analysis of disfluencies", in Proc. of Interspeech 2012, Portland, U.S.A.