An Automatic Prosody Tagger for Spontaneous Speech

Mónica Domínguez
Universitat Pompeu Fabra
C. Roc Boronat, 138
08018, Barcelona, Spain
monica.dominguez@upf.edu

Mireia Farrús
Universitat Pompeu Fabra
C. Roc Boronat, 138
08018, Barcelona, Spain
mireia.farrus@upf.edu

Leo Wanner
ICREA & UPF
C. Roc Boronat, 138
08018, Barcelona, Spain
leo.wanner@upf.edu

Abstract

Speech prosody is known to be central in advanced communication technologies. However, despite the advances of theoretical studies in speech prosody, so far, no large scale prosody annotated resources that would facilitate empirical research and the development of empirical computational approaches are available. This is to a large extent due to the fact that current common prosody annotation conventions offer a descriptive framework of intonation contours and phrasing based on labels. This makes it difficult to reach a satisfactory inter-annotator agreement during the annotation of gold standard annotations and, subsequently, to create consistent large scale annotations. To address this problem, we present an annotation schema for prominence and boundary labeling of prosodic phrases based upon acoustic parameters and a tagger for prosody annotation at the prosodic phrase level. Evaluation proves that inter-annotator agreement reaches satisfactory values, from 0.60 to 0.80 Cohen's kappa, while the prosody tagger achieves acceptable recall and f-measure figures for five spontaneous samples used in the evaluation of monologue and dialogue formats in English and Spanish. The work presented in this paper is a first step towards a semi-automatic acquisition of large corpora for empirical prosodic analysis.

1 Introduction

Speech prosody is known to be central in advanced communication technologies. It is decisive in structuring the message, stressing parts of the message that the interlocutor considers important, and revealing information about the interlocutor's attitude and affection state (Nooteboom, 1997; Wennerstrom, 2001). However, despite the advances of theoretical studies in speech prosody, so far, no sufficiently large, well-annotated prosody material has been created to support empirical studies and drive the research on empirical techniques for analysis and generation of prosodic cues, especially for application in human-computer interaction technologies. Common annotation conventions, such as the ToBI convention (Beckman et al., 2005), provide a descriptive framework of intonation contours and phrasing based upon labels that are language-dependent and rather subjective, which makes it difficult to reach a satisfactory inter-annotator agreement for creating gold standard annotations to train and evaluate algorithms.

It is, therefore, not surprising that empirical research is still based upon rather small laboratory experiments. A further consequence of the lack of sound universal prosody annotation conventions is that current methodologies applied to speech prosody segmentation are still based upon textual and linguistic units (usually words or syntax) rather than on acoustic and phonological units (prosodic phrases and prosodic words). These limitations become an insurmountable barrier for technologies that aim at grasping prosodic cues in spontaneous speech, where many complex prosodic, linguistic and affective phenomena occur (hesitations, incoherent discourse structure, false starts, continuation rising tunes for holding the floor, expression of emotions, speech acts, prosodic disambiguation, etc.). These inherent characteristics of oral language cannot be dealt with using strategies that belong to written language. For instance, sentences with false starts including a filled pause (e.g., *They've never . . . mmm well, my brother's been to Barcelona*).

To overcome the limitations of the current annotation practice and advance in the derivation of more meaningful communicative units from speech as well as in the generation of more natural synthesized speech in the field of human-machine interaction technologies, we need:

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 377–386, Osaka, Japan, December 11-17 2016.

- a parametric language-independent annotation schema of prosody at the acoustic level that can be used by computational models for automatic segmentation and prominence detection;
- prosody taggers and acoustic feature extractors that distill acoustic features from raw speech signals.

In what follows, we address both tasks. First, we present an annotation schema that is implemented as a modular script, deployed as an extended version of the Praat software (Boersma, 2001) into which a functionality for feature annotation and retrieval is incorporated. Such a feature annotation functionality contributes to the independent modular structure and also helps visualization and manual revision of the output within the same Praat environment. Then, we introduce our prosody tagger. Results on inter-annotator agreement and tagger performance compared to a baseline using only F0 cues show that our work is a relevant contribution to the state of the art in the field of speech prosody processing.

The rest of the paper is structured as follows. Section 2 provides an overview of the theoretical approaches in speech prosody, existing automatic tools for labeling prosody and a brief description of the Praat software used for the implementation of our prosody tagger. Section 3 describes the adapted methodology, which is based on theoretical studies of hierarchical prosody and the annotation schema used for manual annotation and for the implementation of the automatic prosody tagger. Section 4 covers the architecture and technical description of the prosody tagger and Praat's extended functionality developed for feature annotation. Section 5 discusses the inter-annotator agreement and evaluation of the performance of our methodology for automatic segmentation and prominence labeling at the prosodic phrase level. Finally, conclusions and future work are drawn in Section 6.

2 Related Work

Theoretical studies of speech prosody have claimed since the 1980's the hierarchical nature of prosodic events in self-contained units (Selkirk, 1984; Ladd, 2008; Gussenhoven, 1984). They describe intonation as a suprasegmental feature, which goes beyond textual segments and has its own structure and phonology. These theoretical studies, especially the one by Selkirk (1984), describe prosody as a hierarchical entity composed of embedded prosodic levels. In our work, we focus on the prosodic phrase (PPh) and the phenomenon of prominence at the PPh level. Our interest in the PPhs stems from the fact that the PPh level has been shown to correlate with the extended thematic structure advocated by Mel'čuk (2001); see, e.g., Domínguez et al. (2014).[1] This correlation proved to be instrumental for the prediction of expressive prosodic contours (Domínguez et al., 2016a).

Recently, a series of models for a computational representation of the intonation contour have been developed and made available in an open initiative for their comparison and further study under the Common Prosody Platform (CPP)[2] (Prom-On et al., 2016): the *Command-Response* (CR) model, the *Autosegmental-Metrical* (AM) model, the *Task-Dynamic* (TD) model and the *Target Approximation* (TA) model. However, these models are limited to a representation of fundamental frequency (F0) contours, which is known to be only one element of speech prosody (Tseng, 2004).

Regarding automatic annotation for prosody involving machine learning techniques, AuToBI[3] (Rosenberg, 2010) was the first publicly available tool to automatically annotate intonation (again, mainly F0 contours and also breaks) with ToBI labels (Silverman et al., 2010). However, AuToBI has several limitations. Thus, it outputs word-by-word annotation, which is a handicap for obtaining a higher-level representation. Furthermore, it is trained on an English corpus of broadcasting radio news, making it domain- and language-specific. In line with AuToBI, ANALOR[4] (Avanzi et al., 2008) is a tool for semi-automatic annotation of French prosodic structure, trained on a small corpus of radio broadcast.

[1] Contrary to other theories of information structure based upon simple theme-rheme division (Steedman,), Mel'čuk (2001)'s extended thematic structure allows embeddedness of thematicity spans and propositions.

[2] http://commonprosodyplatform.org/

[3] http://eniac.cs.qc.cuny.edu/andrew/autobi/

[4] http://www.lattice.cnrs.fr/Analor.html?lang=fr

Like AuToBI, it is domain- and language-specific, but it allows segmentation of an utterance into major prosodic units.

Praat (Boersma, 2001) is an open-source platform for phonetic research widely used by the speech community for annotation, analysis and synthesis purposes. Praat allows the annotation of sound files by means of tiers. Each tier, which is mapped to the whole time-stamp of the associated sound file, includes interval or point annotations that cannot overlap. Each annotation has a label and this label is the only information that can be included into any annotation. Since the labels cannot be extracted as objects in the main Praat window, no action can be scripted based upon smaller units than tiers. Notwithstanding, Praat is a powerful tool, user-friendly, programmable, freely available, running on many platforms, and actively maintained (Mertens, 2004). Due to all these characteristics, a number of Praat-based tools have appeared over the last decade, among them, e.g., ProsodyPro (Xu, 2013). However, many of these tools create a set of parallel tiers, assigning different labels to these tiers, and then, output extracted acoustic features in a text format for further processing using other platforms. For example, Praaline (Christodoulides, 2014) process the *txt* file externally using the R statistical package.

3 Methodology

The main goal of the present work is the development and implementation of a methodology that serves as a scaffolding upon which further improvements and empirical studies can be built upon. One of the requirements, as introduced before, is that this methodology is versatile in the sense that it is language-independent and is able to describe prosodic cues in natural language using a parametric approach. A second goal of our work is to implement a prosody tagger embedded into a Praat-based platform that allows feature annotation and retrieval of any segment below the tier level. In what follows, we introduce the methodology used in the annotation and implementation of a prosody tagger following a discrete representation based upon normalized acoustic parameters (see Section 3.1). The specific annotation guidelines for manual annotation are detailed in Section 3.2. A technical specification of the implementation of the prosodic tagger is outlined in Section 4.

3.1 Acoustic Parameters in Prosodic Units

A key element in our methodology is the concept of prosody as a multidimensional acoustic entity, which involves F0, intensity, and duration acoustic elements. Our methodology involves a combination of normalized acoustic parameters that has been proven by recent studies to yield better results in the task of the prediction of prosodic labels (Domínguez et al., 2016b).

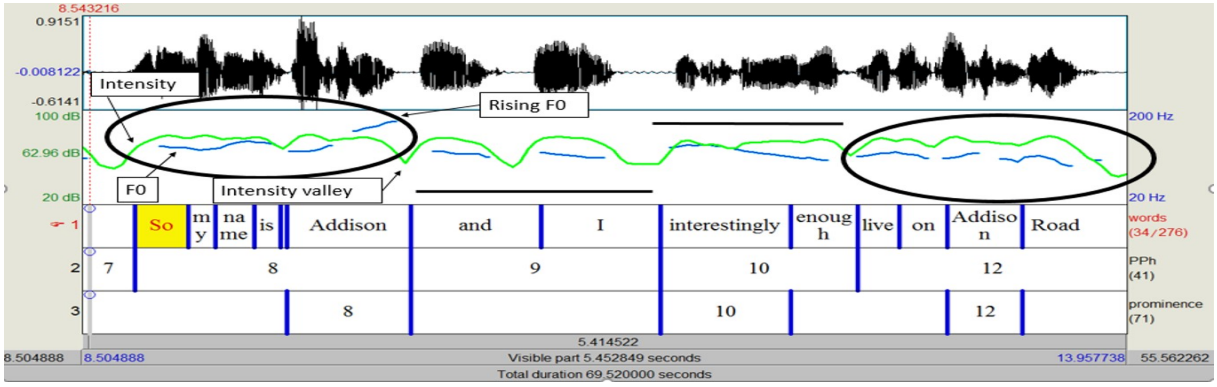

Figure 1: Example of prosodic units.

Figure 1 shows a snapshot of an instance of an utterance containing two propositions (in Mel'čuk (2001)'s terminology). The snapshot corresponds to a continuation rise (L H-H% according to the ToBI convention (Silverman et al., 2010)) in the first proposition and a typical falling tune (H H-L%) in the second. Looking at the graphic representation of the intonation and intensity contours provided by the Praat algorithm, we can clearly observe how these lines form a homogeneous picture (marked by an

ellipse in Figure 1) that corresponds exactly with the PPh division. Nevertheless, as we are dealing with spontaneous speech, there are some areas where the division is not that clear, as can be observed in the central part of the utterance. This homogeneity observed in Figure 1 (which is also auditorily perceived by an expert annotator) can be translated into a vector of normalized acoustic values, which are computed at different levels.

For the representation of the vectors, we adopt a practical approach based upon the actual functions already provided by Praat for speech signal processing. We establish different levels of abstraction, which do not strictly correspond to theoretical representations as such, but follow the idea that prosodic units to be tagged must follow a certain parametric logic in terms of positive or negative deviations at each level in their associated segments. Consequently, we consider as Level 1 (L1) the whole utterance or speech sample; as Level 2 (L2) each voiced segment within the utterance; Level 3 (L3) PPhs once they are segmented. Domínguez et al. (2016b) describe boundary tones in terms of a combination of normalized acoustic parameters, showing also that such a combination performs better than each acoustic element individually for the prediction of prosodic cues.

3.2 Annotation Guidelines

In this section, a set of guidelines for the annotation of prominence and boundaries at the PPh level is introduced. Since the corpus used for evaluation is spontaneous speech, which has inherent difficulties especially for segmentation, specific notes on how to proceed in controversial points are included in these guidelines.

A PPh is defined as a prosodic unit that forms a homogeneous unit in terms of F0 and intensity curves and is signaled by one or a combination of acoustic parameters as outlined in the previous section. A PPh is marked according to the following criteria:

- In case there is one or (usually) a combination of the following conditions: pause, final rising intonation, lengthening of the last word, sharp fall in intensity, a PPh boundary is to be marked.
- In terms of content packaging, a PPh must contain at least one complete concept (usually a predicate with its arguments) of a considerable length relative to the whole utterance and associated voiced segment respectively.
- In spontaneous speech, disfluencies such as disruptions, truncated phrases and hesitations may influence manual labeling of prosodic units. Therefore, all these events are to be included in the closest PPh, unless a pause precedes or follows such disfluencies.
- If the contour following an unvoiced phoneme (with *undefined* F0 value) is perceived as a continuation of the previous F0 contour (forming an homogeneous unit), no boundary is to be inserted. On the contrary, if the F0 contour is significantly different after the F0 phonemic disruption, a boundary is to be marked.

Prominence within each PPh is marked in accordance with the following criteria:

- Prominent words are defined as a combination of one or (usually) several of the following parameters: F0 peak, high intensity, longer duration within its PPh.
- At least one word must be labeled as prominent within each PPh.
- Perceived relevant content must not be used as a criterion to label prosodic prominence (e.g., in noun compounds, an element tends to be perceived more prominent as it carries the semantic meaning of the unit).
- If a combination of acoustic parameters occurs within a word,[5] this word should have more weight than another word showing, for example, an F0 peak and no other acoustic cue.

[5]We refer to textual units in this case, as we are not aiming at segmenting prosodic words yet

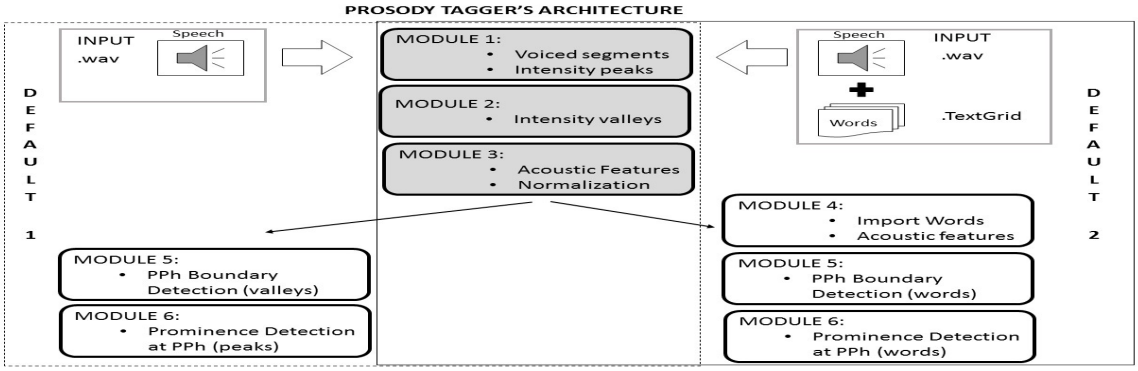

Figure 2: Prosody tagger's architecture.

4 Prosody Tagger Implementation

Our prosody tagger is available as a web service.[6] Any speech sample in *wav* format and associated *TextGrid* with word division can be uploaded, such that the segmentation into PPhs and prominence within the segmented PPhs is displayed on screen and also for download in TextGrid format. All scripts and the extended Praat version for feature annotation are also available under a Creative Common's license[7].

The architecture of the prosody tagger has been conceived as a modular platform such that its optimization and further development (including prosodic word detection) can be attained focusing on specific intermediate steps within the whole pipeline. Acoustic information extracted from different modules is annotated in terms of feature vectors in each segment, including computed z-scores within different prosodic units (so far, from Levels 1 to 3), as introduced in previous sections. Those features can be visualized, retrieved and used for processing at any stage thanks to the extension for feature annotation performed on Praat. Acoustic parameters include, but are not limited to, F0, intensity and duration elements, as Praat allows extraction of a wider range of acoustic parameters (such as jitter, shimmer and pulses, among others).

Figure 2 sketches the modular architecture of the prosody tagger with two possible configurations: (i) Default 1: using only raw audio (as wav file), and (ii) Default 2: using both raw audio (wav file) and importing external word segmentation (in TextGrid format), which must be uploaded by the user. For the Default 2 configuration presented in this study, we have used for word segmentation the proprietary Automatic Speech Recognition system *Scribe*[8] by Vocapia Research.[9] The output of Scribe is converted from xml into TextGrid format. In what follows, a description of each module's functionality is outlined and annotated acoustic features are specified at each stage.

Module 1 uses the wav file and creates a TextGrid using the built-in function in Praat *To TextGrid (silences)*, which automatically detects unvoiced and voiced segments as intervals. Then, a pitch object and an intensity object are extracted from the sound file. The function *To IntensityTier (peaks)* is performed on the intensity object to select salient peaks. The F0 information is extracted at standard frame rates from the pitch object to associate extracted intensity peaks to the ones that involve F0; the distance between these peak candidates is also considered for syllable nuclei detection. A point tier is created and points matching the combination of intensity, F0 and time distance within each voiced segment are annotated. As features, absolute intensity, F0 and the associated voiced interval are stored in each point segment.

Module 2 makes use of the intensity object created in Module 1 to extract intensity valleys using the Praat function *To IntensityTier (valleys)*. Standard intensity frames are selected if their intensity z-score

[6] http://kristina.taln.upf.edu/praatweb/

[7] https://github.com/monikaUPF/modularProsodyTagger

[8] https://scribe.vocapia.com/; Scribe is currently run as a beta version.

[9] http://www.vocapia.com/

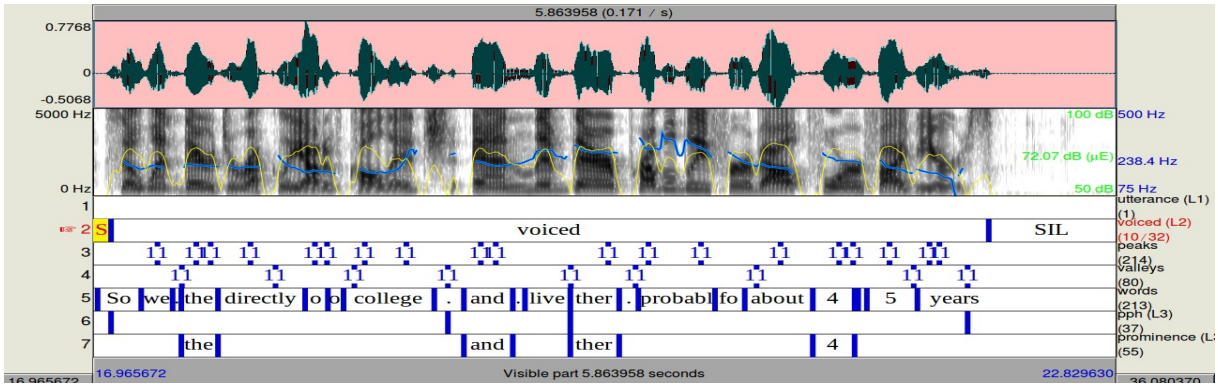

Figure 3: All modules' output for calculation.

(relative to L1) is lower than 0. Then, the lowest values in intensity relative to each voiced fragment (L2) are labeled in a new point tier taking into account the distance between them. Annotated features from this module are: intensity z-score relative to the whole sound (L1) and intensity z-score relative to the associated voiced segment (L2) at each valley point.

Module 3 extracts acoustic values, computes z-scores at available levels, and annotates results as features in each segment. At L1, mean and standard deviation of intensity and F0, together with duration for the whole file, are annotated. These values serve for calculation of z-scores at lower levels in the hierarchy. At L2, annotated features include both absolute values for F0, intensity and duration for further calculation of z-scores in peak and valley tiers (created in Module 1 and 2 respectively) and z-scores derived from L1 values. In the peak and valley tiers, the distance to the previous point is also annotated as a feature. For the first point in the tier, the distance to the boundary of its associated voiced segment is specified as reference.

If a TextGrid with the word segmentation is available, Module 4 exports this tier and annotates features at each marked interval. Consequently, prominence predicted in Module 6 outputs prominent words if these segments are provided by the user. Extracted acoustic parameters and annotated features in this module include: (i) z-scores relative to their associated L2 voiced interval (the z-score values for intensity and F0 are extracted and annotated as features for each word segment obtained by Module 1); (ii) time landmarks, i.e., time of minimum value of intensity and maximum F0 within each word; (iii) duration: absolute duration of the word, and relative duration to the corresponding voiced segment and to the whole sample.

Module 5 uses voiced segments and valleys to predict PPh boundaries. They are derived from the information extracted in the L2 voiced/silence segments detected by Module 1 and from the valleys marked in Module 2. Valleys and peaks contained in each L2 voiced segment are looped in to find the smallest z-score values of intensity within each voiced range and measuring the distance of these valleys to the closest peaks. If the distance of one of the closest peaks is greater than or equal to 0.2 seconds, the z-score is among the minimum in the range, and F0 value is *undefined*, then a PPh boundary is marked.

Finally, Module 6 performs prominence detection on each PPh predicted in the previous module. If no word alignment is available, only syllable peaks predicted in Module 1 are used. Consequently, this module outputs prominent points that correspond to peak points in configuration 1 and prominent word intervals in configuration 2. For calculation of prominence, a combination of F0, intensity and duration cues are taken into account as described in Section 3. Figure 3 displays all tiers created by each module as described above for computation and the final output with the tagging of PPh boundaries and prominent words after the whole pipeline has been executed running a default 2 configuration.

5 Evaluation

A total of five different spontaneous speech samples are used in the evaluation, both for inter-annotator agreement and the prosody tagger's performance: three dialogues in Spanish and two monologues in

American English. Table 1 shows the specific information details for each sample. Dialogues in Spanish are set in a medical context; a male doctor is involved in all of them talking to a patient. Gender is represented in all file names with the convention "f" and "m" for female and male respectively. Files "es_01mm" and "es_02mm" include the same speakers in the same conversational context, where a patient complains to the doctor, but in "es_01mm", the doctor shows a negative response, while in "es_02mm", he acts in a comprehensive and pro-active way. Monologues in English are biographical introductions of the speakers (birthplace, family, recent activities, etc.). Original sound files, transcripts and annotations used in this evaluation will be made available upon request to the authors and acceptance of associated license terms.

		Length	
Filename	Format	Seconds	Words
es_01mm	dialogue	36	196
es_02mm	dialogue	28	150
es_03fm	dialogue	152	545
en_04m	monologue	70	213
en_05f	monologue	30	282
	TOTAL	**316**	**1386**

Table 1: Corpus used in evaluation.

Filename	Prominence	Boundary
es_01mm	0.55	**0.98**
es_02mm	**0.63**	**0.72**
es_03fm	0.51	**0.78**
en_04m	**0.72**	**0.93**
en_05f	**0.69**	**0.70**

Table 2: Inter-annotator agreement Cohen's kappa.

Two expert annotators, proficient in both English and Spanish, have independently labeled both speech samples following the guidelines outlined above in Section 3.2. Cohen's kappa (Cohen, 1960) has been calculated for inter-annotator agreement for each prominence and boundary labeling task. Evaluation is performed on the Default 2 configuration using word segmentation to facilitate the computation and objectiveness of the validation process. A baseline pipeline using only duration and F0 parameters for the same task has been implemented. Inter-annotator kappa results are presented in Section 5.1. The tagger's accuracy, precision and recall compared to the baseline is reported in Section 5.2.

5.1 Inter-annotator agreement

Table 2 provides kappa values for PPh boundary and prominence labeling of our corpus. If annotators label words that are part of the same prosodic word (e.g, they coincide in the final initial word boundary or are separated by a function word, which is usually unstressed), we count this as a partial match for the kappa computation. In order to count matches automatically under Praat, annotators are asked to insert interval boundaries duplicating the word boundaries which are automatically marked, so that we can compare boundary times for the computation of matches.

A kappa within the range of 0.6-0.8 (within a scale between 0 and 1) is considered *satisfactory*, and above 0.8 *perfect* (Cohen, 1960). In Table 2, kappa values that are in line with these thresholds are highlighted in bold for each task (i.e., prominence and boundary labeling within PPh level). Results prove that agreement ranges from 0.51 and 0.98. A higher agreement is observed in the boundary labeling task for all voice samples. No significant differences are observed between English and Spanish samples in boundary detection. However, in prominence labeling, two Spanish samples (files "es_01mm" and "es_-03fm") only reach a *moderate* agreement of 0.55 and 0.51 respectively and, in the overall picture, kappa values for prominence in Spanish are lower than those for English, which might be due to the dialogue format of the samples with shorter interventions, affective states displayed by participants (perceived emotional behavior conveyed by prosody) and quick turn movements between speakers. Nevertheless, we cannot infer that prominence annotation could be language-dependent as such: the corpus we use for evaluation is simply too small for such conclusions. Further research could exploit these techniques, or even a combination of semi-automatic annotation using the prosody tagger presented in this paper, to explore how linguistic parameters such as the discourse type (dialogue in Spanish versus monologue in English), register, gender or speaker idiosyncrasies may affect inter-annotator agreement and tagger's performance in this respect.

5.2 Automatic prosody labeling performance

In order to evaluate the performance of the automatic prosody tagger, we count as full matches those matches that have been labeled as full either by one or by both annotators. For prominence labeling only, we include into the match count also partial matches, i.e., words that coincide in one interval boundary or belong to the same prosodic word – as already done in the inter-annotator agreement exercise. Boundaries that match with a time margin of ± 0.25 seconds are considered to be partial matches. For PPh boundaries, we count it as a match (or true positive), if the automatic tool labels a boundary which has only been labeled by one annotator. Table 3 presents the accuracy, precision, recall and F-measure scores for full matches (F) and full and partial matches (F&P), both for the baseline and our tagger (recall that the baseline uses F0 only).

	Accuracy		Precision		Recall		F-Measure	
	P	B	P	B	P	B	P	B
baseline (F)	0.83	0.89	0.49	0.88	0.22	0.28	0.30	0.42
tagger (F)	**0.84**	0.88	**0.52**	0.58	**0.32**	**0.43**	**0.36**	**0.55**
baseline (F&P)	0.90	0.90	0.84	0.88	0.37	0.28	0.51	0.42
tagger (F&P)	**0.91**	0.89	0.80	0.63	**0.49**	**0.49**	**0.61**	**0.55**

Table 3: Automatic prosody tagger's accuracy, precision and recall.

Table 3 shows that the prosody tagger performs at accuracy rates higher than 0.84 in both prominence (P) and boundary (B) detection tasks. The baseline achieves higher precision figures (especially in boundary detection) than our tagger. A closer look at the output reveals that the baseline marked only those boundaries that included a clear pause, i.e., "safe" candidates. In contrast, the tagger marked not only those clear pauses, but also more subtle boundaries that involved an intensity decrease and not necessarily a pause. On the other side, the tagger reaches considerably higher recall figures than the baseline for both prominence and boundary detection tasks. The F-measure figures show that overall, the tagger performs better. Still, since our methodology is based upon the deviation of normalized values, neutral speech might pose a problem when trying to tag both prominence and boundaries, as there is a tendency towards less variable prosodic cues in this register. Further empirical studies using a semi-automatic approach and optimization of the tagger are needed to have a deeper insight into this and other issues.

6 Conclusions

The integration of a parametric prosody annotation methodology into speech signal processing research is essential in order to reach a close to natural segmentation of spontaneous speech samples and to facilitate the task of the annotation of large corpora for training algorithms for the generation of expressive synthesized speech.

We presented an annotation schema that facilitates a hierarchical acoustic representation of prosody and a tagger that automatically tags speech samples in accordance with this schema. The rather high inter-annotator agreement figures show that our annotation schema is coherent and objective, i.e., does not depend on potentially subjective criteria of the individual annotators. Recall results surpassing the baseline prove that our automatic prosody tagger is flexible enough to support language independent speech signal analysis and detection of prominence and boundaries at the PPh level using a combination of acoustic features, rather than merely F0 contours, as previous empirical and theoretical studies claimed. Improved recall scores also indicate that the number of true positives from the total number of words which actually belong to the positive class, i.e., labeled as positive by the manual annotators, is much higher than the baseline's.

The present implementation may be extended for other applications or further smaller prosodic unit detection (such as prosodic words) due to its modular architecture and open-source philosophy. Moreover, this paper briefly presents an extended Praat functionality for feature annotation to provide easy access and manual revision of the tagger's output as well as retrieval of annotated features at any stage of the computation process. The modularity and prosodically-oriented methodology used in the development of the tagger provides a suitable framework for the deployment of more complex data-driven

approaches, which are able to learn from prosodic units instead of textual units, and thus, get closer to more natural and expressive results when applied to generation of prosodic cues.

All in all, the presented methodology and implementation serves as a platform upon which further research lines and experiments can be run to increase the knowledge in the area of speech technologies and test advanced implementations for human-machine interaction technologies. In the future, we plan to explore the re-implementation of the tagger as a neural network application. Extracted acoustic features combined with linguistic features such as part of speech tag, syntactic dependencies and communicative structure, will be put to the test to observe whether prediction is enhanced, as previously suggested by empirical studies such as (Domínguez et al., 2016b).

Acknowledgements

This work is part of the KRISTINA project, which has received funding from the *European Unions Horizon 2020 Research and Innovation Programme* under the Grant Agreement number H2020-RIA-645012. It has been also partly supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). The second author is partially funded by the Spanish Ministry of Economy and Competitivity through the *Juan de la Cierva* program. The authors also want to acknowledge Yvan Josse and Bianca Vieru from Vocapia Research for their support on speech to text scripts, and Iván Latorre and Joan Codina from Universitat Pompeu Fabra for their support on the implementation of Praat extension for feature annotation and the "Praat on the Web" tool.

References

M. Avanzi, A. Lacheret-Dujour, and B. Victorri. 2008. ANALOR: A tool for semi-automatic annotation of french prosodic structure. In *Proceedings of the 4th International Conference on Speech Prosody*, pages 119–122, Campinas, Brazil.

M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel. 2005. The original ToBI system and the evolution of the ToBI framework. In S A. Jun, editor, *Prosodic Typology – The Phonology of Intonation and Phrasing*.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

G. Christodoulides. 2014. Praaline: Integrating tools for speech corpus research. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. In Sage Publications Inc., editor, *Educational and Psychological Measurement*, pages 37–46.

M. Domínguez, M. Farrús, A. Burga, and L. Wanner. 2014. The Information StructureProsody Language Interface Revisited. In *Proceedings of the 7th International Conference on Speech Prosody*, pages 539–543, Dublin, Ireland.

M. Domínguez, M. Farrús, A. Burga, and L. Wanner. 2016a. Using hierarchical information structure for prosody prediction in content-to-speech applications. In *Proceedings of the 8th International Conference on Speech Prosody*, pages 1019–1023, Boston, USA.

M. Domínguez, M. Farrús, and L. Wanner. 2016b. Combining acoustic and linguistic features in phrase-oriented prosody prediction. In *Proceedings of the 8th International Conference on Speech Prosody*, pages 796–800, Boston, USA.

C. Gussenhoven. 1984. *On the Grammar and Semantics of Sentence Accents*. Foris, Dordrecht.

R. Ladd. 2008. *Intonational Phonology*. Cambridge University Press, Cambridge.

I. A. Mel'čuk. 2001. *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*. Benjamins, Amsterdam, Philadephia.

P. Mertens. 2004. The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of the 2nd International Conference on Speech Prosody*, pages 549–552, Nara, Japan.

S. Nooteboom. 1997. The prosody of speech: Melody and rhythm. In *The Handbook of Phonetic Sciences*. Blackwell Publishers Ltd, Oxford.

S. Prom-On, Y. Xu, W. Gu, A. Arvaniti, H. Nam, and D. H. Whalen. 2016. The common prosody platform (cpp): Where theories of prosody can be directly compared. In *Proceedings of the 8th International Conference on Speech Prosody*, pages 1–5, Boston, USA.

A. Rosenberg. 2010. AutoBI - A tool for automatic ToBI annotation. In *Proceedings of Interspeech*, pages 146–149, Makuhari, Japan.

E. O. Selkirk. 1984. *Phonology and Syntax: The relation between sound and structure*. The MIT Press, Cambridge, Massachussetts.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 2010. ToBI: A standard for labeling English prosody. In *Proceedings of Interspeech*, pages 146–149, Makuhari, Japan.

M. Steedman. Information structure and the syntax-phonology interface. In *Linguistic inquiry*, volume 31, pages 649–689. The MIT Press, Cambridge, Massachussetts.

C. Tseng. 2004. Intensity in relation to prosody organization. In *International Symposium on Chinese Spoken Language Processing*, pages 217–220. IEEE.

A. Wennerstrom. 2001. *The Music of Everyday Speech. Prosody and Discourse Analysis*. Oxford University Press, Oxford.

Y. Xu. 2013. Prosodypro a tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, pages 7–10, Aix-en-Provence, France.