HMM-based Prosodic Structure Model Using Rich Linguistic Context

Nicolas Obin [1], *Xavier Rodet* [1], *Anne Lacheret* [2]

[1] Analysis-Synthesis Team, IRCAM, Paris, France
[2] Modyco Lab., University of Paris Ouest - La Défense, Nanterre, France

nobin@ircam.fr, rodet@ircam.fr, anne.lacheret@u-paris10.fr

Abstract

This paper presents a study on the use of deep syntactical features to improve prosody modeling [1]. A French linguistic processing chain based on linguistic preprocessing, morpho-syntactical labeling, and deep syntactical parsing is used in order to extract syntactical features from an input text. These features are used to define more or less high-level syntactical feature sets. Such feature sets are compared on the basis of a HMM-based prosodic structure model. High-level syntactical features are shown to significantly improve the performance of the model (up to 21% error reduction combined with 19% BIC reduction).

Index Terms: Prosody, Prosodic Structure, Speech Synthesis, High-Level Syntactical Analysis.

1. Introduction

Research on speech synthesis has lead to significant improvements over the past decade that make possible to generate natural speech from text. However, if the synthesized speech sounds *acoustically* natural, it is often considered poor according to the *speaking style* (prosodic artifacts and monotony). Now, modeling the variability in the speaking style (variations of prosodic parameters) is required to provide natural expressive speech in many applications of high-quality speech synthesis such as multi-media (avatar, video game, story telling) and artistic (cinema, theater, music) applications.

In parallel, linguistic studies have investigated phonological models widely in order to formally represent abstract prosodic objects and structure as well as the prosodic / syntactic interface. Phonological models (ToBI for English [1], Prosogram, IntSint, IVTS for French [2, 3, 4]) and expert prosodic predictive models have been proposed ([5, 6, 7] for French). Some attempts have been proposed in order to implement these informations into the automatic speech recognition and synthesis domains: explicit hierarchical prosodic structure modeling for automatic prosodic boundaries detection [8, 9]; prosodic structure predictive models [10, 11]; prosodic structure predictive models from surface syntactic parsing [12, 13]). Recently, robust automatic deep syntactical parsers ([14] for French) have been developed which permit an accurate modeling of the prosodic / syntactic dependencies in a generative framework ([15] for acoustic modeling).

This paper presents a study that aims to model prosodic / syntactic dependencies. It is organized as follows: section 2 presents the linguistic processing chain and the syntactical features extracted from text; section 3 presents the HMM-based model; finally evaluation and results are presented and discussed in sections 4 and 5.

2. High-Level Syntactical Analysis

2.1. Linguistic Processing Chain

An input text (sentence, set of sentences or raw text) is processed by an automatic linguistic parser in order to extract high-level linguistic features (surface and deep syntactical parsing) at the sentence level.

The *Alpage Linguistic Processing Chain* [2] is a full linguistic processing chain for French which is organized as a sequence of processing modules: a *lexer* module (Lefff: a French Morphological and Syntactic Lexicon [16]; SXPipe: a full linguistic preprocessing chain for French [17]), a *parse* module (DyALog: a parser compiler and logic programming environment [18]; FRMG: a FRench Meta Grammar [14]), and a post-processing module.

Deep parsing is performed by the FRMG parser, a symbolic parser based on a compact *Tree Adjoining Grammar* (TAG) for French that is automatically generated from a meta-grammar. The parsing result is then enriched by a series of post-processing modules whose role is to organize all of the information retrieved along the whole linguistic processing.

The output of FRMG is a shared derivation forest that represents all derivation structures that the grammar can build for the input sentence, and indicates which TAG operation (substitution, adjunction, anchoring) took place on a given node of a given tree for a given chunk. This forest is then transformed into a shared dependency forest: anchors of trees related to a given node label are put into a dependency relationship with this label. Node labels are generally associated with their grammatical or syntactical function.

A dependency forest is represented into a *DEP XML* format that incorporates the following items:

- **clusters** that are associated with the forms of the sentence;
- **nodes** that point to a given cluster and are associated to a lemma, a syntactical category and a set of derivations;
- **edges** that connect a source node with a target node are assigned an appropriate label. More precisely, a given edge is associated with a set of *derivations* related to this edge and the related source and target chunk *operations*.

At last the forest is disambiguated by an heuristic-based module that outputs a single dependency tree. In cases where

[1] This study was supported by ANR *Rhapsodie* 07 Corp-030-01; reference prosody corpus of spoken French; French National Research Agency

[2] http://alpage.inria.fr/alpc.en.html

complete parsing could not be achieved, the parser switches from full to partial parsing. This is achieved by a post parsing over partial parses to retrieve the best sets of partial parses covering the input. An example of an output disambiguated dependency graph is shown in figure 2.

2.2. Syntactical Feature Extraction

From the output of the linguistic process described in 2.1, a set of more or less high-level linguistic features is extracted to be used for prosody modeling. The first set of features is related to surface processing while the others are extracted from the deep parsing step.

morpho-syntactical: morphological and syntactical form features such as extracted from the surface processing.

- form segment;
- form lexical category and class (function vs. content form);

form dependency: form dependencies such as extracted in the deep parsing. This set basically encodes the relationship between forms.

- {governor, current, governee} form lexical category and class;
- *edge type and label* between current form and {governor, governee} form;
- *signed dependency distance* between current form and {governor, governee} forms (in forms and in chunks);

recursive chunk: recursive chunks are retrieved in a top-down process according to the operations and associated derivations. For our example sentence (cf. fig. 2), complete recursive chunks are:

(*S* (*AdvP* Longtemps) ((*VP* je me suis couché) (*NP* de bonne heure)))

Recursive chunks are finally transformed into non-recursive chunks by extracting only the leaves of the transformed chunk tree.
The following features are then extracted:

- {governor, current, governee} *chunk category*;
- *edge type and label* between current chunk and {governor, governee} chunks;
- *signed dependency distance* between current chunk and {governor, governee} chunks (in forms and in chunks);
- *chunk depth*;

adjunction: as presented in section 2.1, *adjunctions* represent a specific type of syntactical phenomena. In particular, adjunctions can relate to different text spans (from a single form to a full sentence). Interestingly, adjunction covers a large amount of syntactical phenomena (such as incises, parentheses, subordinate and coordinate clauses, enumerations, ...).

In the FRMG parser formalism, adjunctions can be easily extracted according to specific pattern matching (Fig. 1). Full adjunction is then extracted by retrieving the full depedency descendence from the introducer.
These features are used to extract:

- {governor, introducer, governee} form category;
- *edge type and label* between modified and introducer nodes and between introducer and modifier nodes;
- *signed dependency distance* between the adjunction's introducer and the modified node (in forms and in chunks);

In the case of recursivity, where a given adjunction can be embedded within another adjunction, only the adjunction with the larger span iss extracted.
Syntactical features extracted from text are then used in a prosodic context-dependent model.

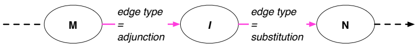

Figure 1: *Generic adjunction pattern. In this figure, M is the governor node, N the governee node, I the introducer.*

3. Prosodic Structure Model

The proposed prosodic structure model is a context-dependent HMM model based on the approach decribed in [10] using a sequential prosodic structure grammar as proposed in [19]. This grammar is based on a hierarchical prosodic description of the concept of prosodic packaging and prosodic prominence. The prosodic grammar is composed of: *major frontier* (FM, frontier of a prosodic group), *minor frontier* (Fm, frontier of an accentual group) and *prosodic prominence* (P, lexical prominence). This grammar is finally transformed into a sequential grammar in order to fullfill the HMM framework.

3.1. Prosodic Structure Model Training

During the training procedure, contextual features are first clustered according to a classification tree estimated according to the minimum entropy criterion [20]. The classification tree is grown using a stop criterion set to 50 observations for a node and then pruned back according to a separate development set. Thus HMM-models $\lambda = \{p(q_0), p(\theta|q), p(q_{n-1}|q_n)\}$ (respectively initial probability, observation probability, and transition probabilities) are estimated for each terminal node of the resulting contextual tree.

3.2. Prosodic Structure Model Prediction

Such models are then used in a HMM inference framework. Let $\mathbf{\Theta} = [\theta_0, ..., \theta_{N-1}]$ be a sequence of contextual observations and $\mathbf{q} = [q_0, ..., q_{N-1}]$ the hidden prosodic structure sequence. Thus,

$$p(\mathbf{q}|\mathbf{\Theta}, \lambda) \propto P(q_0)P(\theta_0|q_0, \lambda) \prod_{n=1}^{N-1} p(\theta_n|q_n, \lambda)p(q_n|q_{n-1})$$

The optimal sequence is estimated according to the maximum likelihood criterion using the Viterbi algorithm.

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}}(p(\mathbf{q}|\mathbf{\Theta}, \lambda))$$

Such a parametric approach appears particularly suitable for prosodic structure modeling since it is possible to estimate speaker-dependent prosodic structure models and thus to model prosodic specific strategies of a given speaker or speaking style (exemple 1, cf. supra).

((tu es bien inhumain$_{F_m}$) (d'avoir perdu ainsi tes enfants$_{F_m}$)$_{F_M}$)
((tu es **bien**$_P$ inhumain$_{F_m}$)$_{F_M}$) ((d'avoir perdu ainsi$_{F_m}$) (tes enfants$_{F_m}$)$_{F_M}$)

Table 1: *Prosodic strategies as infered by speaker-dependent models for the utterance: "Tu es bien inhumain d'avoir perdu ainsi tes enfants !" (A monster you must be to lose your children in this way!). Litlle Tom Thumb, Charles Perrault.*

4. Experiment

4.1. Speech & Text Material

In this study we compared the performance of the proposed prosodic structure model on two very distinct French read-

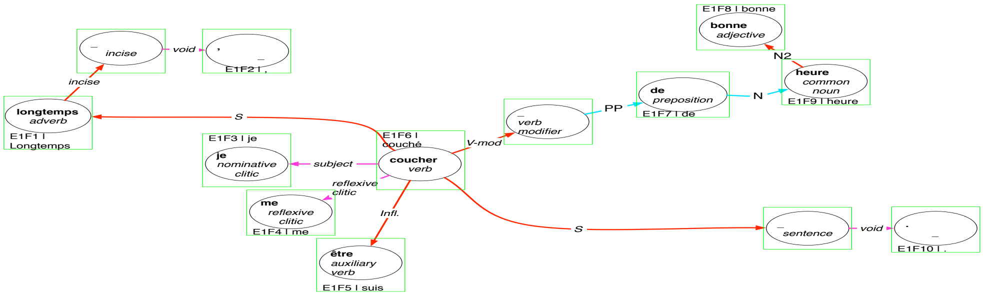

Figure 2: *Disambiguated dependency graph for the sentence: "Longtemps, je me suis couché de bonne heure." ("For a long time I used to go to bed early"). Squares represent clusters (with associated form), circles represent nodes (with associated lemma and lexical category), arrows represents edges (with associated dependency label) going from source node (governor) to target node (governee).*

speech corpora: a *laboratory corpus* with simple linguistic structure and controlled speech (spoken isolate utterances recorded in an anechoic room) and a *multi-media corpus* [3] interpreted by a professional actor. Corpora properties are summarized in table 2:

corpus	speaker gender	speech type	speaker expertise	corpus size	linguistic complexity	prosodic complexity
laboratory	male	read	naive	9h	-	-
multi-media	male	read	actor	5h	+	+

Table 2: *Description of the speech corpora.*

From the comparison of both corpora, it can be clearly expected that the model will drop in performance for the multi-media corpus.

This is due to 1) high linguistic and prosodic complexity: linguistic properties cannot be controlled and professional actors provide a wider variety of prosodic strategies than non-professional (less stereotypical thus less predictable) ; 2) automatic linguistic feature extraction is less robust with highly complex linguistic structures (for instance, complete parsing was achieved for 80% and 52% of the sentences of the laboratory and the multi-media corpus respectively)

Nevertheless this type of corpus presents the advantage of providing rich and various syntactical structures as well as rich prosodic strategies. Such an approach is also justified by the fact that the prosodic model should be robust for any real data as it is required in many multi-media applications.

4.2. Corpus Preprocessing

The following preprocessing chain was applied to the input corpus: phonemic segmentation using *ircamAlign* [21]; syllabification on inter-pausal groups; automatic prosodic frontiers detection with *Analor* [19]; automatic syllable-based prominence detection with *ircamProm* [22].

4.3. Prosodic Structure Model's Parameters

Different sets of linguistic features distributed on a more or less high-level feature scale were defined:

- **morpho-syntactical** (linguistic units: form + syllable-based baseline features: syllabic phonological features (phonemic content and syllabic structure));
- **dependency** (linguistic unit: form);
- **chunk** (linguistic unit: chunk);
- **adjunction** (linguistic unit: adjunction);

For each feature set, low-level linguistic features were computed on each linguistic unit with a first-order left-to-right context: *locational* and *weight* features (position and number of a given unit within higher level units).

4.4. Evaluation scheme

We compared syllable-based sequential models trained with the different linguistic feature sets, each feature set being added to the previous ones in the training process accordingly to the proposed scale. Models were evaluated within a 10-folder cross validation framework. Two measures were used to evaluate models' performance:

Bayesian Information Criterion [23]: a normalized likelihood measure that is used in particular for model selection. Models BIC were estimated on the training set;

Weighted Cohen's Kappa [24]: provides a paired agreement measure in the case of ordinal categorical rating, where categorical labels are ordered along a continuous scale. Kappa measures provide statistical agreement measures which account for that expected by chance. In particular, weighted Cohen's Kappa penalizes errors according to the nature of the disagreed labels [4]. Linear Cohen's Kappa was used in this experiment on the evaluation set.

5. Results & Discussion

Figure 3(a & b) presents the mean performance measures obtained for the laboratory and multi-media corpora.

In both cases, performance increases as higher-level feature sets are added. This improvement is particularly significant for the chunk and adjunction feature sets (for the adjunction feature set: 21% and 11% of Kappa reduction; 19% and 7.5% of BIC reduction were observed on the laboratory and multi-media corpora respectively when compared to the initial feature set). Conversely, there is no significant difference between the form-based feature sets (morpho-syntactical and form dependencies). These results suggest that prosodic structure is more closely related to large syntactical units rather than form unit only.

When comparing the performance obtained for each corpus, there is a clear drop in performance for the multi-media corpus.

[3] audio-book: "*Du côté de Chez Swann*", first volume of "*A la Recherche du Temps Perdu*" from french writer Marcel Proust

[4] for instance: a confusion on the presence of a frontier (FM or Fm vs. P or NP) is more important that a confusion on the precise type of a frontier (FM vs. Fm)

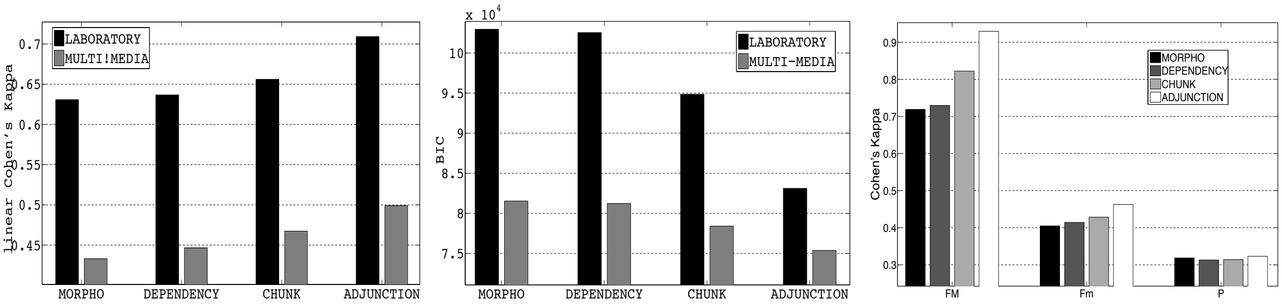

Figure 3: *(a) Overall Linear Cohen's Kappa according to the different feature sets. (b) Bayesian Information Criteria value according to the different feature sets. (c) Cohen's Kappa for each prosodic label according to the different feature sets.*

This confirms the expected tendency as discussed in section 4.1. Secondly, if the performance tendency related to the feature sets is still observed, this improvement is much lower than it is for the laboratory corpus. These results should be related to the fact that 1) the actor provides more varied and complex prosodic strategies; 2) the automatic feature extraction is less robust thus less reliable on complex syntactical structures.

Investigating the performance in finer detail reveals that the performance is clearly dependent on the prosodic label (fig. 3 c).

Frontier prediction presents substantial (FM) and moderate (Fm) performance while lexical prominence (P) prediction, only fair performance. This is consistent with performances found in the litterature for other prosodic structure systems. Secondly, the performance gain does not uniformly affect the different prosodic labels. This improvement is clearly significant for the prosodic frontiers prediction, especially for the major frontiers, when there is no improvement for the lexcial prominence prediction. Such results confirm a significant relationship between prosodic packaging and syntactical structures and a poor relationship between lexical prominence and syntactical structures. Since lexical prominence encodes lexical phenomena which are strongly related to semantic and discursive linguistic levels, this hardly appears predictable from a syntactical description only. Higher-level linguistic features are thus needed to accurately model the location of such prominences.

6. Conclusion

We have presented a prosodic structure model based on the automatic extraction of rich linguistic context. High-level syntactical features have been shown to significantly improve the performance of the prosodic model. In particular, syntactical features such as chunk and adjunction features reveal a substantial relationship with the prosodic structure. This confirms existing evidence for linguistic study carried on the syntactic-prosodic interface. However, syntactical features failed to accurately model lexical prominence. Further research will focus on the typology of the model: on one hand, by estimating the prosodic structure model's parameters in a unified HMM framework and on the other by explicitely modeling the hierarchical nature of the prosodic structure. This will be done within a hierarchical HMM or more generally within a WTA (Weighted Tree Automata) framework. Finally, other linguistic levels, for example semantic, will be introduced in order to improve lexical prominence modeling.

7. References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: A stan-dard for labeling english prosody," in *ICSLP*, 1992, pp. 867–870.

[2] P. Mertens, "The prosogram : Semi-automatic transcription of prosody based on a tonal perception model," in *Speech Prosody*, Nara, Japan, 2004, pp. 549–552.

[3] D. Hirst, A. Di Cristo, and R. Espresser, *Prosody: Theory and Experiments*. M. Horne, 2000, ch. Levels of representation and levels of analysis for the description of intonation systems.

[4] B. Post, *Tonal and phrasal structures in French intonation*. Holland Academic Graphics, 2000.

[5] F. Dell, "L'accentuation dans les phrases en français," *Forme sonore du langage: structure des représentation en phonologie*, pp. 65–122, 1984.

[6] E. Delais-Roussarie, "Vers une nouvelle approche de la structure prosodique," *Langue Française*, vol. 126, 2000.

[7] P. Mertens, "Quelques allers-retours entre la prosodie et son traitement automatique," *Le français moderne*, vol. 72, no. 1, pp. 39–57, 2004.

[8] N. Segal and K. Bartkova, "Prosodic structure representation for boundary detection in spontaneous speech," in *ICPhS*, 2007, pp. 1197–1200.

[9] J. Teppereman and S. Narayanan, "Tree grammars as models of prosodic structure," in *Interspeech*, 2008, pp. 2286–2289.

[10] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech and Langage*, vol. 10, pp. 155–185, 1996.

[11] K. Syrdal, A., J. Hirschberg, J. McGory, and M. Beckman, "Automatic tobi prediction and alignment to speed manual labeling of prosody," *Speech Communication*, vol. 33, no. 1-2, pp. 135–151, 2001.

[12] A. Black and P. Taylor, "Assigning intonation elements and prosodic phrasing for english speech synthesis from high level linguistic input," in *ICSLP*, 1994, pp. 715–718.

[13] V. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE transactions on ASLP*, vol. 16, no. 4, pp. 797–811, 2008.

[14] E. Villemonte de La Clergerie, "From metagrammars to factorized TAG/TIG parsers," in *IWPT*, Vancouver, Canada, Oct. 2005, pp. 190–191.

[15] N. Obin, P. Lanchantin, M. Avanzi, Lacheret-Dujour, and X. Rodet, "Toward improved hmm-based speech synthesis using high-level syntactical features," in *Speech Prosody*, 2010.

[16] B. Sagot, L. Clément, E. Villemonte de La Clergerie, and P. Boullier, "The lefff 2 syntactic lexicon for french: architecture, acquisition, use," in *LREC*, Genova, Switzerland, 2006.

[17] B. Sagot and P. Boullier, "From raw corpus to word lattices: robust preparsing processing," in *L&TC 2005*, 2005, p. 2005.

[18] E. Villemonte de La Clergerie, "Dyalog: a tabular logic programming based environment for nlp," in *CSLP*, Barcelona, Spain, 2005.

[19] M. Avanzi, A. Lacheret-Dujour, and B. Victorri, "Analor: A tool for semi-automatic annotation of french prosodic structure," in *Speech Prosody*, 2008, pp. 119–122.

[20] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Wadsworth & Brooks, 1984.

[21] P. Lanchantin, A. Morris, X. Rodet, and C. Veaux, "Automatic phoneme segmentation with relaxed textual constraints," in *LREC*, Marrakech, Morroco, 2008.

[22] N. Obin, X. Rodet, and A. Lacheret-Dujour, "Prominence model: a probabilistic framework," in *ICASSP*, Las Vegas, U.S.A, 2008.

[23] Y. Bishop, J. Fienberg, and P. Holland, *Discrete Multivariate. Analysis*. MIT Press, 1975.

[24] J. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and Psychological Measurement*, vol. 33, pp. 613–619, 1973.