HMM-based Prosodic Structure Model Using Rich Linguistic Context
Nicolas Obin, Xavier Rodet, Anne Lacheret

HAL Id: hal-00589069
https://hal.archives-ouvertes.fr/hal-00589069
Submitted on 27 Apr 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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, morphosyntactical 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] 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 [6, 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 1 is a full linguistic processing chain for French which is organized as a sequence of processing modules: a lexer module (Leff: 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, ad- junction, 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

http://alpage.inria.fr/alpc.en.html

---

1This study was supported by ANR Rhapsodie 07 Corp-030-01; reference prosody corpus of spoken French: French National Research Agency
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 de-
pendency graph is shown in figure [2].

2.2. Syntactical Feature Extraction

From the output of the linguistic process described in [2], 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 fea-
tures 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 
depth parsing. This set basically encodes the relationship be-
tween forms.

- \{governor, current, governee\} form lexical category and 
class;
- \textit{edge type and label} between current form and \{governor, 
governee\} form;
- \textit{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:

\begin{align*}
(S(AdvP Longtemps) \cdot (VP je me suis couché) \cdot (NP de bonne heure)))
\end{align*}

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

adjunction: as presented in section 2.1 adjunctions represent 
a specific type of syntactical phenomena. In particular, adjunc-
tions 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, subor-
dinate and coordinate clauses, enumerations, ...).

In the FRMG parser formalism, adjunctions can be easily 
easily extracted according to specific pattern matching (Fig. 1). 
Complete adjunctions are then extracted by retrieving the full de-
pendency descendance from the introducer.

These features are used to extract:

- \{governor, introducer, governee\} form category;
- \textit{edge type and label} between modified and introducer nodes 
and between introducer and modifier nodes;
- \textit{signed dependency distance} between the adjunction’s intro-
ducer 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 is extracted.

Syntactical features extracted from text are then used in a 
prosodic context-dependent model.

2.3. Prosodic Structure Model

The proposed prosodic structure model is a context-dependent 
HMM model based on the approach described in [10] using a se-
quential 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 accent-
ual group) and prosodic prominence (P, lexical prominence). 
This grammar is finally transformed into a sequential grammar 
in order to fulfill the HMM framework.

3.1. Prosodic Structure Model Training

During the training procedure, contextual features are first clus-
tered 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 \( \lambda = \{p(q_0), p(q|\theta), p(q_n|q_{n-1})\} \) (respec-
tively initial probability, observation probability, and transition 
probabilities) are estimated for each terminal node of the result-
ing contextual tree.

3.2. Prosodic Structure Model Prediction

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

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

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

\[
\hat{q} = \arg \max_q (p(q|\Theta, \lambda))
\]

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 

Table 1: Prosodic strategies as inferred 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!). Little 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-
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 interpreted by a professional actor. Corpora properties are summarized in table 2.

<table>
<thead>
<tr>
<th>corpus</th>
<th>speaker gender</th>
<th>speech type</th>
<th>speaker expertise</th>
<th>corpus size</th>
<th>linguistic complexity</th>
<th>prosodic complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>laboratory</td>
<td>male</td>
<td>read</td>
<td>native</td>
<td>1h</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>multi-media</td>
<td>male</td>
<td>read</td>
<td>native</td>
<td>1h</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

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.


4for 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)
This confirms the expected tendency discussed in section 6.1. 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. 6c). 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 literature 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 lexical 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 explicitly 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