Reformulating Prosodic Break Model into Segmental HMMs and Information Fusion

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

[1] Sound Analysis and Synthesis, IRCAM, Paris, France
[2] Modyco Lab., University of Paris Ouest - La Défense, Nanterre, France
nobin@ircam.fr, lanchantin@ircam.fr

Abstract

In this paper, a method for prosodic break modelling based on segmental-HMMs and Dempster-Shafer fusion for speech synthesis is presented, and the relative importance of linguistic and metric constraints in prosodic break modelling is assessed [1]. A context-dependent segmental-HMM is used to explicitly model the linguistic and the metric constraints. Dempster-Shafer fusion is used to balance the relative importance of the linguistic and the metric constraints into the segmental-HMM. A linguistic processing chain based on surface and deep syntactic parsing is additionally used to extract linguistic informations of different nature. An objective evaluation proved evidence that the optimal combination of the linguistic and the metric constraints significantly outperforms both the conventional HMM (linguistic information only) and segmental-HMM (equal balance of linguistic and metric constraints), and confirmed that the linguistic constraint is prior to the metric.

Index Terms: speech prosody, prosodic break, segmental-HMM, Dempster-Shafer fusion.

1. Introduction

Linguistic studies generally assume that the production of a prosodic punctuation marker - *prosodic break* - results from the integration of various and potentially conflictual constraints, in particular the *syntactic* and the *metric* constraints [1, 2, 3, 4]. A prosodic break is primarily produced by speakers and can be used by listeners to clarify the structure of the utterance. Simultaneously, secondary cognitive constraints (performance constraints) tend to produce a segmentation into prosodic breaks with an optimal configuration [5], in particular with respect to the metric regularity [2]. These constraints conflict in the production of a prosodic structure, and secondary extra-linguistic constraints often override the primary linguistic constraint.

In speech synthesis, the adequate insertion of prosodic breaks guarantees the intelligibility, the naturalness, and the variety of the synthesized speech. Statistical methods have been proposed to combine linguistic and metric constraints based on segmental models [6, 7, 8]) in the modelling and adaptation of prosodic breaks. However, the proposed methods remain generally based on surface syntactic informations (POS) solely, while deep syntactic informations are ignored. Additionally, the relative importance of the linguistic and the metric constraints is not considered, or inadequately formulated.

In this study, a statistical method to combine linguistic and metric constraints in the modelling of prosodic breaks is proposed based on segmental HMMs and Dempster-Shafer fusion, and the relative importance of the linguistic and the metric constraints is assessed depending on the nature of the linguistic information. A discrete segmental HMM is used in which prosodic breaks are modelled conditionally to the linguistic context in which they are observed, and the distance across successive prosodic breaks (length of a prosodic phrase) is explicitly modelled. Dempster-Shafer fusion is additionally employed to balance the relative importance of the linguistic constraint and the metric constraint into the segmental HMM. Segmental HMMs are objectively evaluated with respect to different sets of linguistic contexts, and the relative importance of the linguistic and the metric constraints is assessed.

This paper is organized as follows: segmental HMMs and their application to prosodic break modelling are presented in section 2, Dempster-Shafer fusion is presented in section 3. The evaluation is described and discussed in sections 4 and 5.

2. Segmental HMMs

Segmental HMMs [9, 10, 11, 12] were introduced in speech recognition in which state sequences are explicitly represented as *segments* with an explicit modelling of the segment state-occupancy. Segmental HMM is a generalization of hidden Markov model (HMM) that addresses two principal limitations of the conventional hidden Markov model: 1) state duration modelling, and 2) assumption of conditional independence of the observation sequence given the state sequence.

The reformulation of prosodic break modelling into a segment model requires to reformulate prosodic breaks as segments. Actually, a *prosodic break* instantiates a prosodic segment (*prosodic phrase*) that is defined as the segment left/right bounded by a prosodic break. Thus, the modelling of prosodic breaks can reformulated in terms of prosodic segments.

Let define $\mathbf{q} = [\mathbf{q}_1, \ldots, \mathbf{q}_T]$ the sequence of linguistic contexts of length T, where $\mathbf{q}_t = [q_t(1), \ldots, q_t(L)]^\top$ is the (Lx1) linguistic context vector which describes the linguistic characteristics associated with the t-th syllable, $\mathbf{l} = [l_1, \ldots, l_T]$ the corresponding sequence of prosodic labels, where l_t denotes the prosodic label associated with the t-th syllable,

[1] This study was partially funded by “La Fondation Des Treilles”, and supported by ANR Rhapsodie 07 Corp-030-01; reference prosody corpus of spoken French; French National Agency of research; 2008-2012.

Copyright © 2011 ISCA

$\mathbf{s} = [s_1, ...s_K]$ the associated sequence of prosodic phrases of length K, and $\mathbf{d} = [d_1, ...d_K]$ the corresponding segment state-durations, where d_k denotes the length of prosodic phrase s_k.

In prosodic break modelling, the segment model can be simplified as follows:

1. one segment: $s_k = [\ l_{[t_{k-1}+1:t_k-1]} = \bar{b},\ l_{t_k} = b\ ]$

2. segment transition = 1

where: $\mathbf{t} = [t_1, \ldots, t_K]$ denotes the sequence of segment boundaries, and b denotes a prosodic break and $\bar{b}$ the absence of a prosodic break.

2.1. Parameters Estimation

During the training, the estimation of the context-dependent segmental HMM parameters is simplified, and the parameters of the linguistic model $\boldsymbol{\lambda}^{(\text{linguistic})}$ and segment duration model $\boldsymbol{\lambda}^{(\text{metric})}$ are estimated separately.

$$\boldsymbol{\lambda} = \left(\boldsymbol{\lambda}^{(\text{linguistic})}, \boldsymbol{\lambda}^{(\text{metric})}\right) \tag{1}$$

The linguistic model $\boldsymbol{\lambda}^{(\text{linguistic})}$ is estimated using the context-dependent discrete HMM described in [13]. First, linguistic contexts are clustered so as to derive a context-dependent tree. Then, a context-dependent HMM $\boldsymbol{\lambda}^{(\text{linguistic})} = (\lambda_{S_1}^{(\text{linguistic})}, \ldots, \lambda_{S_M}^{(\text{linguistic})})$ is constructed from the set of terminal contexts $S = (S_1, \ldots, S_M)$ of the decision-tree, where λ_{S_m} denotes the HMM parameters associated with the context S_m.

The segment duration model $\boldsymbol{\lambda}^{(\text{metric})}$ is estimated with a normal distribution.

2.2. Parameters Inference

During the synthesis, the segment sequence $\widehat{(\mathbf{s}, \mathbf{d})}$ is determined so as to maximize the conditional probability of the segment sequence $\mathbf{s}$ and the segment duration sequence $\mathbf{d}$ given the linguistic context sequence $\mathbf{q}$:

$$\widehat{(\mathbf{s}, \mathbf{d})} = \underset{\mathbf{s}, \mathbf{d}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{s}, \mathbf{d} | \mathbf{q}) \tag{2}$$

The determination of the segment sequence $\widehat{(\mathbf{s}, \mathbf{d})}$ can be proved to be equivalent to the determination of the prosodic break sequence $\widehat{\mathbf{l}}$ as follows:

$$\widehat{\mathbf{l}} = \underset{\mathbf{l}}{\operatorname{argmax}} \prod_{k=1}^{K} \frac{\mathrm{p}(\mathbf{l}_{[t-d_k+1:t-1]} = \bar{b}, l_t = b\ | \mathbf{q}_{[t-d_k+1:t]})}{\mathrm{p}(\mathbf{l}_{[t-d_k+1:t-1]} = \bar{b}, l_t = b)} \times \mathrm{p}(d_k | \mathbf{l}_{[t-d_k+1:t-1]} = \bar{b}, l_t = b) \tag{3}$$

$$= \underset{\mathbf{l}}{\operatorname{argmax}} \prod_{k=1}^{K} \underbrace{\mathrm{p}_o(t_k)}_{\text{observation probability}} \quad \underbrace{\mathrm{p}_s(t_k)}_{\text{segment probability}} \tag{4}$$

where $\mathrm{p}_s(l_{t_k}) = \mathrm{p}(d_k |\ \mathbf{l}_{[t-d_k:t-1]} = \bar{b}, l_t = b)$ denotes the partial probability that the k-th segment with duration d_k ends at time t_k, and $\mathrm{p}_o(l_{t_k}) \propto \mathrm{p}(\mathbf{l}_{[t-d_k+1:t]} = \bar{b}, l_{t_k} = b\ | \mathbf{q}_{[t-d_k+1:t]})$ the partial observation probability over the k-th segment with duration d_k.

The solution to this problem is achieved with a reformulation of the conventional *Viterbi Algorithm* (VA) for segmental HMMs [12].

3. Dempster-Shafer Fusion

In the formulation of segmental HMMs, the segment probability and the observation probability are equally considered. However, linguistic studies pointed out that the linguistic and the metric constraints are not of equal importance in the production of a prosodic break. In particular, the metric constraint is generally assumed to be secondary compared to the linguistic constraint.

Dempster-Shafer theory [14] is a mathematical theory commonly used for information fusion in statistical processing. In particular, Dempster Shafer theory provides a proper probabilistic formulation for information fusion, in which the *reliability* that can be conferred to different sources of information can be explicitly formulated. In the Dempster-Shafer fusion, PDFs can be reformulated into mass functions (MFs) to account for the reliability that can be conferred to each PDF, and then combined with the Dempster-Shafer fusion rule. Mass functions are defined on $\mathcal{P}(\Omega)$, where Ω denotes the state alphabet, and $\mathcal{P}(\Omega)$ the total set of combinations of Ω.

In order to balance the relative importance of the linguistic constraint $\mathrm{p}_o(l_t)$ and the metric constraint $\mathrm{p}_s(l_t)$ into the segmental HMM, one of the PDFs is alternatively replaced by a mass function (MF), while the other remains a PDF:

$$m_o(l_t) = \alpha\ \mathrm{p}_o(l_t) \qquad m_o(\Omega) = 1 - \alpha \tag{5}$$

$$m_s(l_t) = \beta\ \mathrm{p}_s(l_t) \qquad m_s(\Omega) = 1 - \beta \tag{6}$$

where α and β denote the reliability that is associated with the observation probability $\mathrm{p}_o(l_t)$ and the segment probability $\mathrm{p}_s(l_t)$ respectively, and $m_o(\Omega)$ and $m_s(\Omega)$ the model *ignorance*.

The Dempster-Shafer fusion of m_o and m_s is then given by:

$$\begin{aligned} (m_o \oplus m_s)(l_t) &\propto \alpha(1-\beta)\ \mathrm{p}_o(l_t) \\ &+ \alpha\beta\ \mathrm{p}_o(l_t)\ \mathrm{p}_s(l_t) \\ &+ \beta(1-\alpha)\ \ \mathrm{p}_s(l_t) \end{aligned} \tag{7}$$

Hence,

$$(m_1 \oplus m_2)(l_t) \propto \begin{cases} \mathrm{p}_o(l_t), & \alpha = 1,\ \beta = 0 \quad \textcircled{1} \\ \mathrm{p}_s(l_t), & \alpha = 0,\ \beta = 1 \quad \textcircled{2} \\ \mathrm{p}_o(l_t)\ \mathrm{p}_s(l_t), & \alpha = 1,\ \beta = 1 \quad \textcircled{3} \end{cases}$$

① denotes that the observation probability is considered only (conventional HMM), ② denotes that the segment probability is considered only, and ③ denotes that the segment and observation probabilities are equally considered (conventional segmental HMM). In the latter case, the expression is equivalent to the conventional Bayes combination rule.

Finally, the relative confidence α and β are rewritten into a single weight (α, β) so that the relative importance of the linguistic and the segment probabilities is linearly interpolated from the metric constraint solely to the linguistic constraint solely. Thus: $(\alpha, \beta) = -1$ will refer to $\alpha = 0$ and $\beta = 1$, $(\alpha, \beta) = 0$ to $\alpha = 1$ and $\beta = 1$, and $(\alpha, \beta) = +1$ to $\alpha = 1$ and $\beta = 0$.

4. Evaluation

The evaluation was conducted to assess the relative importance the linguistic and the metric constraints, and their combination in prosodic break modelling. In particular, a large range of combination of linguistic contexts was used to estimate context-dependent segmental HMMs, and various combinations of the linguistic and metric constraints were compared. Two baseline models were used for the comparison: the conventional punctuation rule-based model (P) in which a prosodic break is inserted after each punctuation marker, and the segmental HMM estimated with the conventional morpho-syntactic linguistic context (M).

4.1. Evaluation scheme

The comparison of context-dependent segmental HMMs was conducted using different set of linguistic contexts and different combination of the linguistic and the metric constraints. Evaluation was conducted according to a 10-fold cross-validation. F-measure was used as performance measure and a paired Student t-test [15] was employed to assess whether a significant difference exists between the models being compared.

4.2. Speech Material

Two French read-speech databases were compared: a *laboratory* and a *multi-media* corpus. The laboratory corpus is composed of short sentences that were selected in order to design a phonetically well-balanced speech database. Each sentence was separately read by a non-professional French speaker (9 hours). The multi-media corpus is the novel "*Du côté de Chez Swann*" (*"Swann's Way"*) by the French writer Marcel Proust. The text was read by a professional actor in the context of an audio-book format (7 hours). The laboratory corpus consists of simple linguistic structures and controlled prosody, while the multi-media corpus consists of complex linguistic structures and a rich prosody variety.

4.3. Linguistic Contexts

Linguistic informations were extracted from text using a French linguistic processing chain that includes surface and deep syntactic parsing [16, 17]. The extracted syntactic features were classified into different sets according to the nature of the syntactic information. Morpho-syntactic (M) informations correspond to the conventional syntactic information (POS) used in speech prosody modelling. Dependency (D) and constituency (C) were compared in their relevancy in speech prosody modelling. Additionally, adjunction (A) covers a large variety of syntactic constructions potentially related to speech prosody (e.g., relative clause, incise) and was introduced for comparison. Finally, segmental-HMMs were compared with respect to any combination of the different linguistic feature sets. A detailed description of the linguistic contexts used is presented in [18, 13].

5. Results & Discussion

Table 1 summarizes the mean performance obtained for the laboratory and multi-media speech databases depending on the linguistic feature set, and the comparison of the metric model only, the conventional segmental-HMM, the linguistic model only, and the optimal configuration of linguistic and metric constraints into the segmental-HMM.

The optimal configuration significantly outperforms the conventional segmental-HMM and the linguistic model for

context	$\mathrm{p}_{\mathrm{opt}}$	α, β	p_s	$\mathrm{p}_s\mathrm{p}_o$	t-test	p_o	t-test
laboratory							
MDCA	**96.3**	+0.44	65.4	92.1	*<0.001*	95.0	*<0.001*
MCA	**96.0**	+0.48	65.4	92.1	*<0.001*	94.7	*<0.001*
CA	**96.0**	+0.54	65.4	92.0	*<0.001*	94.6	*<0.001*
DCA	**95.8**	+0.56	65.4	91.7	*<0.001*	94.6	*<0.001*
DA	**94.1**	+0.41	65.4	89.1	*<0.001*	92.6	*<0.001*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
M	**78.3**	+0.23	65.4	75.5	*<0.001*	74.2	*<0.001*
P	**66.3**	-	-	-	-	-	-
multi-media							
MDCA	**75.3**	+0.70	39.0	70.0	*<0.001*	74.0	*0.03*
MCA	**75.2**	+0.58	39.0	69.6	*<0.001*	73.6	*0.04*
DCA	**74.2**	+0.68	39.0	68.6	*<0.001*	72.8	*0.02*
CA	**73.7**	+0.73	39.0	67.4	*<0.001*	72.6	*0.2*
MC	**69.6**	+0.65	39.0	65.5	*<0.001*	68.0	*0.1*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
M	**59.2**	+0.62	39.0	56.7	*0.03*	58.7	*0.5*
P	**55.1**	-	-	-	-	-	-

Table 1: Ranked F-measure for the optimal configuration, the metric model only p_s, the conventional segmental-HMM $\mathrm{p}_s\mathrm{p}_o$, and the linguistic model only p_o. Significance test for the comparison of the optimal configuration, the conventional segmental-HMM and the linguistic model.

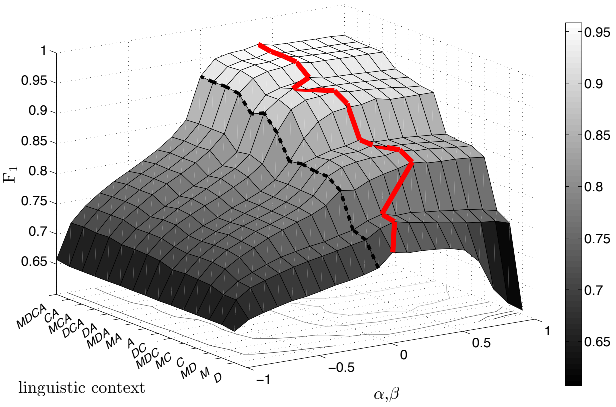

Figure 1: Laboratory corpus: F_1 measure depending on the linguistic context and the balance α,β of metric and linguistic constraints. The dotted line denotes the segmental-HMM, and the red line the optimal combination of the metric and linguistic constraints.

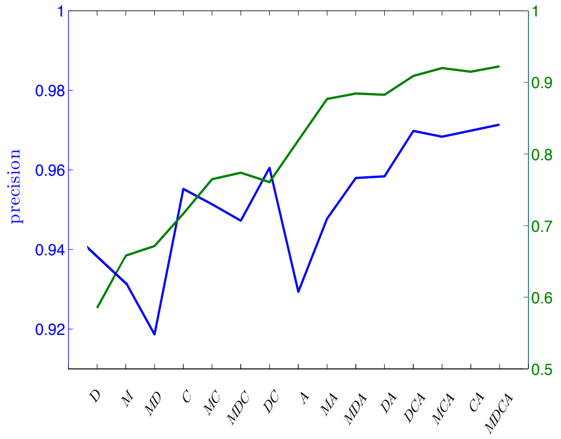

Figure 2: Laboratory corpus: precision and recall of the optimal model depending on the linguistic context.

all of the linguistic feature sets, and corresponds to a prior importance of the linguistic constraint over the metric constraint. The balance of the linguistic and the metric constraints varies depending on the relevancy of the linguistic or/and the metric constraint. In particular, the optimal configuration gradually tends to the linguistic constraint when the linguistic information increase in reliability ((α, β) varies from +0.23 to +0.56 for the laboratory corpus in correlation with a linguistic performance which varies from 74.2% to 95%), and is very close to the linguistic constraint when the metric constraint is not reliable ((α, β) varies from +0.58 to +0.73 in correlation with a metric performance of 39%).

The relevancy of the different linguistic features in prosodic break modelling (figure 1) confirms and refines observations reported in [13]. Adjunction (A) and to a less extent constituency (C) are the most relevant single linguistic contexts (91.7% and 83.8% for the laboratory corpus, 63.2% and 65.3% for the multi-media corpus), and their combination (CA) is strongly relevant (96.0% and 73.7% for the laboratory and the multi-media speech databases respectively). Morpho-syntactic (M) and dependency (D) are slightly relevant linguistic contexts (78.3% and 73.8% for the laboratory corpus, 59.2% and 52.0% for the multi-media corpus). The optimal performance is obtained with the combination of all of the linguistic contexts (MDCA) (96.3% and 75.3% for the laboratory and multi-media speech databases respectively).

Additionally, recall and precision mutually increase when the linguistic description is enriched (figure 2). However, the increase in recall is large (from 62% to 92% for the laboratory corpus, and from 42% to 67% for the multi-media corpus) compared to the increase in precision (from 92% to 97% for the laboratory corpus, and from 76% to 82% for the multi-media corpus). Thus, the enrichment of the linguistic description significantly decreases the omission of prosodic breaks, while the false insertion of prosodic breaks remains globally marginal regardless to the linguistic description.

The increase in performance obtained with the enrichment of the linguistic description is significantly larger compared to that obtained with the integration of the metric constraint. For the linguistic constraint, the increase in performance is of 18% by comparison of the conventional morpho-syntactic context (78.3%) and the optimal linguistic context (96.3%). For the combination of the linguistic and the metric constraints, the increase in performance does not exceed 4% (74.2% and 78.3% for the conventional morpho-syntactic context), and 2% with a rich linguistic description (95% and 93.3% for the optimal linguistic context).

Finally, the performance significantly varies depending on the speech database. The overall performance, and the increase in performance due to the enrichment of the linguistic description and the combination of the linguistic and the metric constraints are significant larger for the laboratory corpus compared to the multi-media corpus. The difference may be simply interpreted in terms of the reliability of the syntactic analysis, and eventually by the difference in prosodic variety of the speakers.

6. Conclusion

In this paper, a statistical method that combines linguistic and metric constraints in the modelling of prosodic breaks was proposed based on segmental HMMs and Dempster-Shafer fusion, and the relative importance of the linguistic and the metric constraints was assessed depending on the nature of the linguistic informations. The optimal combination of the linguistic and the metric constraints into segmental-HMM was proved to significantly outperform the conventional segmental-HMM and the linguistic model only. The linguistic constraint was shown to be prior to the metric constraint, and the optimal configuration to gradually tend to the linguistic constraint when the linguistic description is enriched, or when the metric constraint is slightly reliable. Finally, the increase in performance obtained by the integration of the metric constraint and its combination with the linguistic constraint remains slight compared to that obtained with the enrichment of the linguistic description. In further studies, the segment model will be refined to improve the modelling of the metric constraint and its combination with the linguistic constraint, and will be evaluated for the modelling of various speaking styles.

7. References

[1] E. Selrik, *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge: MIT Press, 1984.

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

[3] G. Bailly, “Integration of rhythmic and syntactic constraints in a model of generation of French prosody,” *Speech Communication*, vol. 2, pp. 137–146, 1989.

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

[5] J. Gee and F. Grosjean, “Performance structures: a psycholinguistic and linguistic appraisal,” *Cognitive Psychology*, vol. 15, pp. 411–458, 1983.

[6] M. Ostendorf and N. Veilleux, “A hierarchical stochastic model for automatic prediction of prosodic boundary location,” *Journal of Computational Linguistics*, vol. 20, no. 1, pp. 27–54, 1994.

[7] H. Schmid and M. Atterer, “New statistical methods for phrase break prediction,” in *International Conference On Computational Linguistics*, Geneva, Switzerland, 2004, pp. 659–665.

[8] P. Bell, T. Burrows, and P. Taylor, “Adaptation of prosodic phrasing models,” in *Speech Prosody*, Dresden, Germany, 2006.

[9] M. Russel and R. Moore, “Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition,” in *International Conference on Acoustic, Speech, and Signal Processing*, Tempa, USA, 1985, pp. 2376–2379.

[10] S. Levinson, “Continuously variable duration hidden Markov models for automatic speech recognition,” *Computer Speech and Language*, vol. 1, pp. 29–45, 1986.

[11] M. Gales and S. Young, “Segmental HMMs for speech recognition,” in *European Conference on Speech Communication and Technology*, Berlin, Germany, 1993, pp. 1579–1582.

[12] M. Ostendorf, V. Digalakis, and O. Kimball, “From hmm’s to segment models: a unified view of stochastic modeling for speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[13] N. Obin, A. Lacheret, and X. Rodet, “HMM-based prosodic structure model using rich linguistic context,” in *Interspeech*, Makuhari, Japan, 2010, pp. 1133–1136.

[14] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[15] G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters*. John Wiley Sons, 1978.

[16] B. Sagot, “The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French,” in *International Conference on Language Ressources and Evaluation*, Valletta, Malte, 2010, pp. 2744–2751.

[17] E. Villemonte de La Clergerie, “From metagrammars to factorized TAG/TIG parsers,” in *International Workshop On Parsing Technology*, Vancouver, Canada, Oct. 2005, pp. 190–191.

[18] N. Obin, P. Lanchantin, A. Lacheret, and X. Rodet, “Towards improved HMM-based speech synthesis using high-level syntactical features,” in *Speech Prosody*, Chicago, U.S.A., 2010.