A Multi-Level Context-Dependent Prosodic Model Applied to Durational Modeling

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

[1]Analysis-Synthesis Team, IRCAM, Paris, France
[2]Modyco Lab., University of Paris-La Défense, Nanterres, France
nobin@ircam.fr, rodet@ircam.fr, anne@lacheret.com

Abstract

We present in this article a multi-level prosodic model based on the estimation of prosodic parameters on a set of well defined linguistic units. Different linguistic units are used to represent different scales of prosodic variations (local and global forms) and thus to estimate the linguistic factors that can explain the variations of prosodic parameters independently on each level. This model is applied to the modeling of syllable-based durational parameters on two read speech corpora - laboratory and acted speech. Compared to a syllable-based baseline model, the proposed approach improves performance in terms of the temporal organization of the predicted durations (correlation score) and reduces model's complexity, when showing comparable performance in terms of relative prediction error.
Index Terms : speech synthesis, prosody, multi-level model, context-dependent model.

1. Introduction

Research on speech synthesis has brought significant improvements over the past decade that makes possible to generate natural speech from text. However, if the synthesized speech sounds *acoustically* natural, it is often considered poor accordingly to the *way of speaking* (prosodic artefacts and monotony). Now modeling the variability in the way of speaking (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 (theater, music) applications. Despite growing attention to prosody modeling over the past few years, one of the major drawback of actual prosody models remains the monotony of the generated prosodic parameters. That problem (averaging problem) appears mainly due to two inter-related causes : a lack of linguistic knowledge extracted from the text that could be used to explain more accurately the observed prosodic variations and a representation of the prosodic parameters which is essentially a superposition of acoustic forms observed on a set of different + / - well defined linguistic units related to different communication functions.

A short overview of linguistic units and associated +/- linguistic parameters affecting these units leads to distinguish at least three main classes of prosodic scales and associated linguistic parameters :
1) *global variations* : overall properties of prosodic parameters (mean, variance). This level is associated with global characteristics of a speaker and to speaking styles associated with specific discourse genres.
2) *local variations* : variations on the smallest linguistic units (sub-state of a phoneme, syllable or phoneme). These variations are associated with phonological properties of these levels (co-articulation, syllabic structure, accentuation).
3) *intermediate variations* : variations on a set of units larger than the syllable and + / - linguistically well defined (accentual group, interpausal group, prosodic group, intonational phrase, period, verbal construction, discourse sequence, ...) and associated with + / - linguistic factors : physiological (f_0 declination), modalities (questions, ...), syntactical (prosodic contrasts related to some specific syntactical sequence), semantic (informational structure) and discursive.

In order to model the variability of these parameters, it is necessary to determine an appropriate representation of prosodic parameters and then to extract and estimate the effects of high-level linguistic features (syntactic, semantic, discursive) on the observed prosodic parameters. At the signal level several approaches have been proposed to represent the variations of the fundamental frequency, HMM phone state-based [1], syllable-based [2], a set of well defined linguistic units [3], or a set of units that are estimated with unsupervised statistical methods [4, 5]. For duration, a phone-based [6, 7, 8] or a syllable-based [9, 10] representation. If the linguistic studies define the syllable or syllable-like [11] unit as the minimal unit of prosodic rhythm, some studies have shown that rythm variations were also used as prosodic cues by speakers on larger units such as speech rate variations on some specific verbal constructions (*oral parenthesis* [12]) as well as informational and discourse structure [13]. Thus in the similar manner as for fundamental frequency variations, it could be usefull to represent durational variations on different linguistic units, thus to estimate the linguistic factors that affect these variations on each unit.

We propose in this article a syllable-based duration model based on multi-level context-dependent analysis. This paper is organized as follow : in section 2 our proposed approach is presented ; in section 3 speech material and evaluation scheme are presented ; results are discussed in section 4.

2. A Multi-Level Context-Dependent Model

Let $L = \{l_1, l_2, ..., l_N\}$ a set of continuous and non-recursive linguistic units (i.e. phonem, syllable, prosodic group, period, ...), $\theta = \{\theta_{l_1}, \theta_{l_2}, ..., \theta_{l_N}\}$ the durational features related to the linguistic levels l_i and $F = \{F_{l_1}, F_{l_2}, ..., F_{l_N}\}$ a set of linguistic features that describe linguistic level l_i.

Accordingly to the General Superpositional Model [14], the duration $\theta(n)$ of the minimal linguistic unit l_1 could be written as :

$$\theta(n) = \bigoplus_{l \in L} \bigoplus_{k \in l} \theta_{l,k}(n)$$

For each linguistic level l_i, a model typology of θ_{l_i} conditionally to F_{l_i} is estimated using the Minimum Description Length

(MDL) criterion with normal distributions [15].

Then given a sequence of linguistic features F_{l_i}, the optimal sequence of observations $\hat{\Theta}_{l_i}$ is estimated as the one that maximizes the conditional probability of observations Θ_{l_i} conditionnaly to F_{l_i} :

$$\hat{\Theta}_{l_i} = \underset{\Theta_{l_i}}{argmax} P(\Theta_{l_i}|F_{l_i})$$

which is according to the independance hypothesis the sequence of means :

$$\hat{\Theta}_{l_i} == [\mu_{l_i,1}; ...; \mu_{l_i,K}]$$

where :

$$\mu_{l_i,k} = \underset{\theta_{l_i,k}}{argmax} P(\theta_{l_i,k}|F_{l_i,k})$$

In contrast to models based on modeling durational features on a single linguistic unit (phoneme, syllable), the proposed approach shows several advantages :
1) distinguishing several linguistic units in the representation of durational features variations enables to explicite the superposition of prosodic forms jointly observed on a given unit,
2) each prosodic level (speech rate, duration syllabic residual, ...) can be modeled and controlled independently from each other,
3) estimate the set of linguistic parameters affecting each linguistic unit independently.

3. Evaluation

3.1. Material

The proposed model has been evaluated on two read speech +/- neutral described in table 1.

type	gender	size	syllable duration mean [std]
laboratory speech	male	> 9 hrs	220 [130] ms.
acted speech (dubling)	male	> 5 hrs	160 [100] ms.

TAB. 1 – Description of the used corpora.

All analysis were conducted within the *ircamCorpusTools* [16] framework. Corpora were segmented using *ircamAlign* [17], syllabified on each interpausal group with *Liaphon* [18] and syllable prominences have been automatically estimated using *ircamProm* [19].

3.2. Model's parameters

3.2.1. Prosodic features and Linguistic units

The proposed model in this study is a syllable-based duration model. In this experiment durational features were chosen as being speech rate on interpausal group and residual syllable duration.

3.2.2. Linguistic factors

In this experiment low-levels linguistic parameters were used, such as locational features (position of a given unit within higher level units), weight features (number of observations of a given linguistic unit within higher level units) and phonological features (syllabic structure and prominence). The used linguistic parameters on each linguistic unit are presented in table 2.

Syllable level

Phonological features

- {preceding, current, succeeding} syllable type,
- {preceding, current, succeeding} syllable structure sequence (onset, nucleus, coda),
- phone label of {preceding, current, succeeding} syllable nucleus,
- phone class in {preceding, current, succeeding} syllable nucleus,
- phone label sequence in {preceding, current, succeeding} syllable,
- phone class sequence in {preceding, current, succeeding} syllable,
- phonem sequence in {preceding, current, succeeding} syllable onset,
- phonem sequence in {preceding, current, succeeding} syllable coda,
- prominence state of {preceding, current, succeeding} syllable.

Locational features

- syllable ordinal position in current prosodic group,
- syllable ordinal position in sentence,
- current prosodic group ordinal position in sentence,
- syllable caterogical position (begin ; middle ; end) in current prosodic group,
- syllable categorical position (begin ; middle ; end) in sentence,
- current prosodic group categorical position (begin ; middle ; end) in sentence,

Weight features

- phone number in {preceding, current, succeeding} syllable,
- phone number in {preceding, current, succeeding} syllable onset,
- phone number in {preceding, current, succeeding} syllable coda,
- syllable number in {preceding, current, succeeding} prosodic group,
- syllable number in {preceding, current, succeeding} sentence,

Interpausal group level

Locational features

- current prosodic group ordinal position in sentence,
- current prosodic group categorical position (begin ; middle ; end) in sentence.

Weigth features

- number of syllables in {previous ; current ; next} prosodic group,
- number of syllables in sentence,
- number of prosodic groups in sentence,

TAB. 2 – List of the contextual features used for each linguistic level.

3.3. Evaluation scheme

We have compared our approach to a baseline syllable-based model in which all described linguistic features were projected on the syllable unit.
Speaker-dependent models were estimated on a training corpus of variable size and performances were estimated on a test corpus of fixed size (30mn.)
In order to evaluate the performance as a function of the amount of training data, variable size training sets were used according to the following sizes : {1mn ; 2mn ; 5mn ; 30mn ; 1hr ; 2hrs ; 5hrs ; [9hrs]}.
Performance measures were chosen as being the relative prediction error and correlation score between observed and predicted syllable durations. Relative error measure removes the influence of observed durations on the error measure (observed syllable durations varie from 14 to 900 ms. for the laboratory speech corpus and from 17 to 910 ms. for the acted speech corpus -

pauses included). Correlation score is a performance parameter used to estimate the goodness of the predicted temporal structure. This last parameters appears more adequate for measuring prosodic prediction performance since prosody is more related to temporal variations rather than absolute error measurement.

4. Results and Discussion

Results are presented in figure 1. An overall comparison of the obtained performances on both corpus shows better performances (relative errors and correlation scores) for the laboratory speech than for the acted speech in all cases. Even if these corpora shows comparable overall relative dispersion (table 1), such difference could be explained by the fact that laboratory speech is more prototypical than acted speech and then more predictable from linguistic features.

Our proposed approach improves performance in terms of correlation score in almost all cases (with exception of the laboratory corpus with a 1mn. training size). Performance improvement is more significant on the acted corpus for which performance gain varies from 6 to 20% compared to the baseline model. However the proposed approach shows a little degradation in relative errors performance (worse for the laboratory speech and similar for the acted speech corpus)

Gain in correlation score could be explained by several properties of our proposed model :
1) modeling durational features on different linguistic units enables to explicite the observed acoustic forms on each level separately, thus removing from the syllabic level variations due to speech rate variations. This leads to a reduction of the dispersion of residual syllable durations.
2) the correlation score is little sensitive to offset errors caused by speech rate prediction errors.
In contrast, the performance degradation in terms of relative error of prediction is due both to the fact that speech rate prediction errors cause a prediction bias on all syllables of the considered interpausal group, but also by the lack of linguistic features that can explain such speech rate variations.

If we look at the evolution of the performances regardless of the model used, we see a clear asymptotic behavior and the gain in performance is no more significant after 1or 2 hours training size. However, this asymptotic behavior in terms of performance is observed jointly to a significant increase in terms of complexity of the models (table 3). On the one hand this means that if low-level linguistic features could be used to estimate robust models with a relatively small amount of data, these remain insufficient to explain more accurately the observed variations. The increase in complexity jointly observed could be explained by the fact that increasement of the number of observations increases at the same time the number of competitive linguistic parameters at each step of the estimattion of the typology of the model. This tends to increase errors on the prioritization of these parameters. Then a larger number of linguistic parameters is needed to obtain similar results.

In terms of complexity (table 3), the proposed model shows a significant reduction in the complexity of the models obtained when a sufficient amount of data is available to learn robust models. (small reduction observed for the laboratory speech from 5hrs. and a significant reduction for the acted speech from 1hrs.). This is mainly due to the fact that the baseline model needs more linguistic parameters to explain variations in speech rate, especially when the model is mixed by trying to jointly explain syllable duration and speech rate variations.

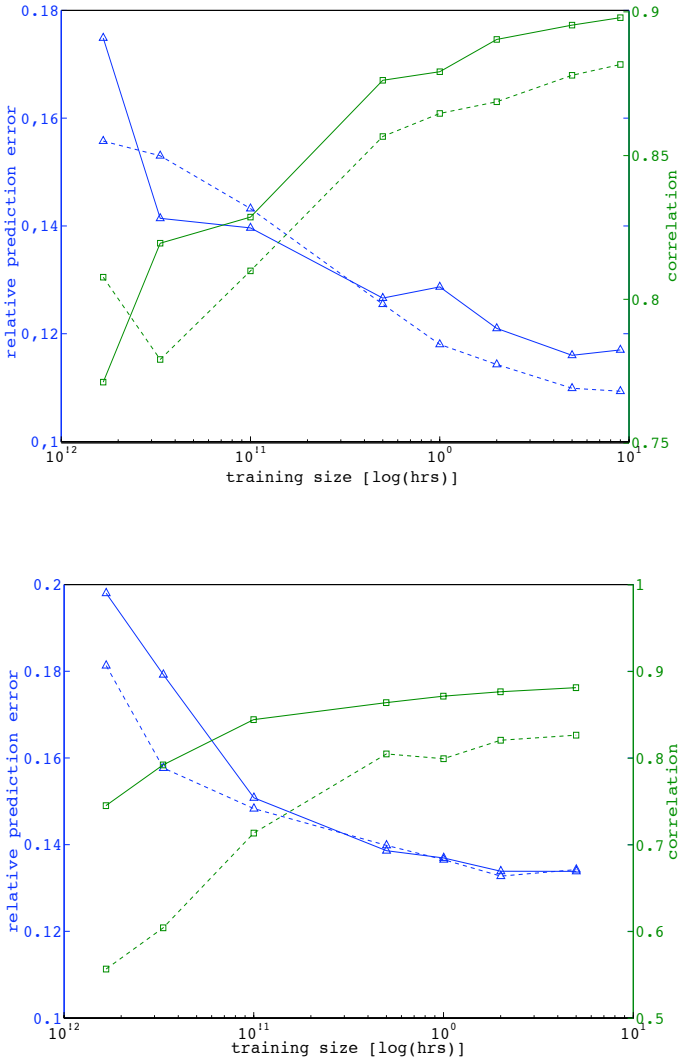

FIG. 1 – Performance as a function of the training set size. Top : laboratory speech. Bottom : acted speech. Baseline model is represented with dash-lines. Proposed model is represented with plain-lines. Relative prediction errors are plotted with triangles and correlation score with squares.

	laboratory speech		**acted speech**	
	syllable	multi-level	syllable	multi-level
1 mn.	16	29	19	24
2mn.	31	36	28	35
10mn.	39	48	41	30
30mn.	103	126	95	89
1h.	151	167	143	116
2h.	236	250	230	192
5h.	488	463	425	331
9h.	766	722	NA	NA

TAB. 3 – Models complexity as a function of the training set size.

5. Conclusion

A multi-level approach of prosodic parameters variations applied to durational features has been presented. This approach explicitly represents the different observation levels of prosodic forms and estimate the linguistic parameters that affects these forms separately on each linguistic unit. Compared to a baseline syllable duration model, the proposed approach improves the performance of the predicted temporal structure (correlation score) and reduces the complexity of the models while providing comparable performance in terms of relative prediction error .

In further works, we will focus on two aspects of prosody modeling. At the acoustic level, we will explicitely model the prosodic temporal structure by introducing dependencies between acoustic observations.
At the symbolic level, we will extract high-level linguistic features (morpho-syntactic, syntactic, semantic, discursive) from text that will be used to improve the estimation of speech rate variations over intermediate linguistic units.

6. Acknowledgements

This study was supported by :

- ANR Rhapsodie 07 Corp-030-01 ; reference prosody corpus of spoken French ; French National Agency of research (ANR) ; 2008-2012.,
- Programmes Exploratoires Pluridisciplinaires (PEPS), CNRS/ST2I, 2008-2010.

7. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," in *Proc. of Eurospeech*, Paris, France, 1999, pp. 2347–2350.

[2] C. Boidin and O. Boëffard, "Generating intonation from a mixed cart-hmm model for speech synthesis," in *Proc. of Interspeech*, Brisbane, Australia, 2008.

[3] J. Latorre and M. Akamine, "Multilevel parametric-base f0 model for speech synthesis," in *Interspeech*, Brisbane, Australia, 2008.

[4] Y. Morlec, "Génération multiparamétrique de la prosodie du français par apprentissage automatique," PhD. Thesis, INPG, Grenoble, 1997.

[5] B. Holm, "Sfc : un modèle de superposition de contours multiparamétriques pour la génération automatique de la prosodie - apprentissage automatique et application à l'énonciation de formules mathématiques," PhD. Thesis, INPG, Grenoble, 2003.

[6] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-markov model based speech synthesis," in *Proc. ICSLP*, 2004, pp. 1397–1400.

[7] B. Gao, Y. Qian, Z. Wu, and F. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood," in *Proc. of Interspeech*, Brisbane, Australia, 2008.

[8] J. Yamagishi, H. Kawai, and T. Kobayashi, "Phone duration modeling using gradient tree boosting," *Speech Communication*, vol. 50, no. 5, pp. 405–415, 2008.

[9] S.-H. Chen, W.-H. Lai, and Y.-R. Wang, "A new duration modeling approach for mandarin speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 308–320, 2003.

[10] K. Sreenivasa Rao and B. Yegnanarayana, "Modeling durations of syllables using neural networks," *Computer Speech and Language*, vol. 21, no. 2, pp. 282–295, 2007.

[11] P. Barbosa and G. Bailly, "Generating segmental duration by p-centers," in *Proc. of the Fourth Workshop on Rhythm Perception and Production*, Bourges, France, 1996, pp. 163–168.

[12] F. Gachet and M. Avanzi, "Les parenthèses en français : Etude prosodique," *Verbum*, 2009.

[13] F. Koopmans-van Beinum and M. van Donzel, "Relationship between discourse structure and dynamic speech rate," in *Proc. of ICSLP*, 1996.

[14] T. Mishra, J. Van Santen, , and E. Klabbers, "Decomposition of pitch curves in the general superpositional intonation model," in *Speech Prosody*, Dresden, Germany, 2006.

[15] J. O'Dell, "The use of context in large vocabulary speech recognition," PhD. Thesis, Cambridge University, 1995.

[16] C. Veaux, B. Beller, D. Schwarz, and X. Rodet, "Ircamcorpustools : an extensible plateform for speech corpora exploitation," in *Proc. of ELREC*, Marrakech, Morroco, 2008.

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

[18] F. Béchet, "Liaphon : un système complet de phonétisation de textes," *Traitement Automatique des Langues*, pp. 47–67., 2008.

[19] N. Obin, X. Rodet, and A. Lacheret-Dujour, "A syllable-based prominence model based on discriminant analysis and context-dependency," in *Proc. of SPECOM*, St-Petersbug, Russia, 2009.