Stylization and Trajectory Modelling of Short and Long Term Speech Prosody Variations

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

[1] Analysis-Synthesis Team, IRCAM, Paris, France
[2] Modyco Lab., University of Paris Ouest - La Défense, Nanterre, France
nobin@ircam.fr, anne.lacheret@u-paris10.fr, rodet@ircam.fr

Abstract

In this paper, a unified trajectory model based on the stylization and the modelling of f_0 variations simultaneously over various temporal domains is proposed[1]. The syllable is used as the minimal temporal domain for the description of speech prosody, and short-term and long-term f_0 variations are stylized and modelled simultaneously over various temporal domains. During the training, a context-dependent model is estimated according to the joint stylized f_0 contours over the syllable and a set of long-term temporal domains. During the synthesis, f_0 variations are determined using the long-term variations as trajectory constraints. In a subjective evaluation in speech synthesis, the stylization and trajectory modelling of short and long term speech prosody variations is shown to consistently model speech prosody and to outperform the conventional short-term modelling.

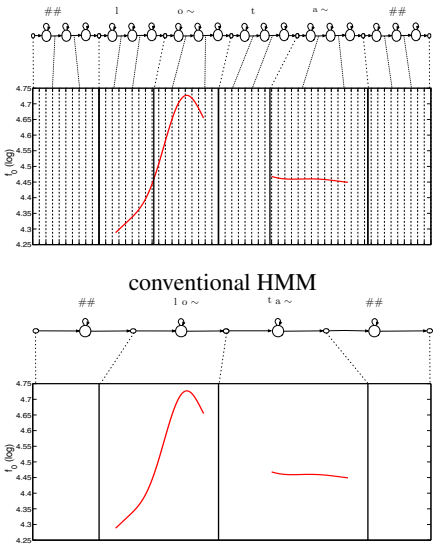
Index Terms: speech prosody, stylization, trajectory model, speech synthesis.

1. Introduction

In parallel to the development of high-quality speech synthesis systems [1], the modelling of speech prosody has raised as a major concern to improve the naturalness, the liveliness, and the variety of the synthetic speech. Speech prosody is generally described as the co-occurrence of acoustic gestures occurring simultaneously over different temporal domains [2, 3] and associated to different communicative functions (linguistic, expressive). A high-quality modelling of speech prosody is desirable for natural and expressive speech synthesis and adequate modelling of speaking style, and a prerequisite in real multi-media applications (e.g., avatars, story telling, dialogue systems, numeric arts).

A variety of methods has been proposed to model speech prosody variations (f_0 [4], temporal structure [5]), and local and global variations [6, 7]. However, conventional methods usually models short-term variations of speech prosody (*frame-based*, or *instantaneous variations*), while long-term variations of speech prosody are not explicitly considered. Recent studies have been proposed to integrate long-term variations into HMM modelling, either for the modelling of f_0 variations [8, 9], or with extension to state-duration modelling [10]. However, the proposed methods remain a *mixed model*, i.e. the conventional model is used to model the *instantaneous* variations of f_0, while stylization of long-term variations are used as trajectory constraints only. In particular, the *instantaneous* variations remain the minimal and target temporal domain for the modelling of speech prosody.

In this paper, a *unified* trajectory model based on the stylization and the joint modelling of f_0 variations over various temporal domains is proposed. In the proposed approach, the syllable is used as the minimal temporal domain for the description of speech prosody, and f_0 variations are stylized and modelled simultaneously over various temporal domains which cover short-term and long-term variations. During the training, a context-dependent model is estimated according to the joint stylized f_0 contours over the syllable and a set of long-term temporal domains. During the synthesis, f_0 variations are determined using the long-term variations as trajectory constraints.



syllable-based HMM with stylization of f_0 contours

Figure 1: Schematic comparison of *frame-based* and *syllable-based* modelling of f_0 variations.

2. Stylization of Speech Prosody

The *Discrete Cosine Transform* (DCT) is used to stylize the f_0 variations over various temporal domains [11] (figure 2). The

[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.

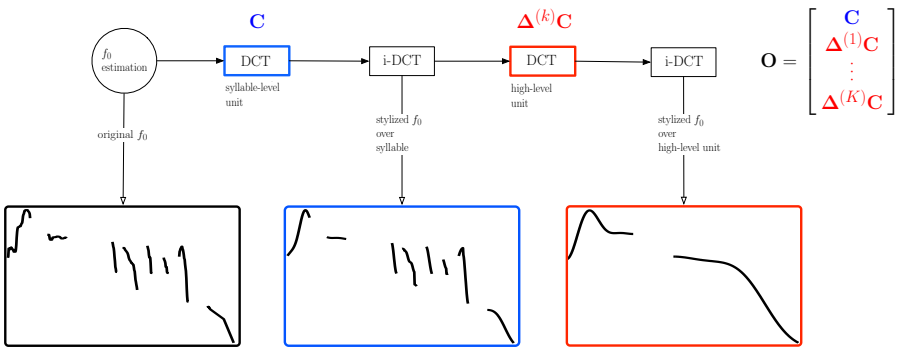


Figure 2: Instantaneous estimation of f_0, short-term stylization over syllable, and long-term stylization over prosodic group.

principle of the DCT is to decompose f_0 contours on a basis of slowly time-varying functions defined by zero-phase cosine functions $\boldsymbol{\phi} = (\cos(\omega_1), \ldots, \cos(\omega_T))$ at discrete frequencies $\omega_k = \frac{\pi}{2T}(2k+1)$, where T is the length of the temporal domain used for stylization.
The stylized f_0 contour is then obtained by inverse transform of the K order truncated DCT ($K \leq T$):

$$f_0(t) = \sum_{k=1}^{K} \alpha_k c_k \cos(\omega_k t) \quad (1)$$

where c_k is the k-th term of the DCT, and α_k a term used for normalization.

Two classes of temporal domains are defined for the stylization of f_0 variations:

Syllable context accounts for f_0 variations occurring on the syllable and its immediate context (0-order represents the f_0 variations over the syllable, 1-order the f_0 variations over the 1-left-to-right syllable context, ...);

Linguistic contexts account for f_0 variations occurring on long-term prosodic units (e.g., minor and major prosodic groups). A minor prosodic group is defined as the prosodic unit that ends with an intermediate prosodic boundary, and is used for rhythmic grouping typical of French. A major prosodic group is defined as the prosodic unit that ends with a major prosodic boundary.

F_0 variations are stylized using a 5-order DCT. F_0 is linearly interpolated in the logarithmic domain prior to the stylization. The stylization over various temporal scales aims at representing f_0 variations with more or less details, and to model short and long term dependencies.

3. Trajectory Model

The *Trajectory Model* has been introduced in HMM-based speech synthesis to explicitly model the dynamic (local variations) of the speech parameters [6]. In this study, syllable is assumed as the minimal temporal domain for the description of speech prosody, and f_0 variations are stylized and modelled simultaneously over different temporal domains: short-term variations correspond to the stylization of f_0 contours over the syllable, and long-term variations correspond to the stylization of f_0 contours over long-term temporal domains. During the training, a context-dependent HMM is estimated from the joint short-term and long-term variations. During the synthesis, the short-term variations are determined so as to maximize the conditional probability of the short-term variations under the constraint of the long-term trajectories.

3.1. Parameters Estimation

Let $\mathbf{q} = [\mathbf{q}_1, \ldots, \mathbf{q}_N]$ be the sequence of linguistic contexts, where $\mathbf{q}_n = [q_n(1), \ldots, q_n(L)]^\top$ is a (Lx1) linguistic vector which describes the linguistic characteristics associated with the n-th syllable.
Let $\mathbf{c} = [\mathbf{c}_1, \ldots, \mathbf{c}_N]$ be the static observation sequence of stylized f_0 contours over the syllable-level unit, where $\mathbf{c}_n = [c_n(1), \ldots, c_n(D)]^\top$ is a (Dx1) observation vector which describes the short-term f_0 characteristics associated with the n-th syllable.
Let $\boldsymbol{\Delta}^{(k)}\mathbf{c} = [\boldsymbol{\Delta}^{(k)}\mathbf{c}_1, \ldots, \boldsymbol{\Delta}^{(k)}\mathbf{c}_N]$ be the dynamic observation sequence of stylized f_0 contours over the k-th long-term temporal domain, where $\boldsymbol{\Delta}^{(k)}\mathbf{c}_n = [\Delta^{(k)}c_n(1), \ldots, \Delta^{(k)}c_n(D)]^\top$ is a (Dx1) observation vector which describes the long-term f_0 characteristics associated with the n-th syllable.
Let $\mathbf{o} = [\mathbf{o}_1, \ldots, \mathbf{o}_N]$ be the augmented observation sequence, where $\mathbf{o}_n = [\mathbf{c}_n^\top, \boldsymbol{\Delta}^{(1)}\mathbf{c}_n^\top, \ldots, \boldsymbol{\Delta}^{(K)}\mathbf{c}_n^\top]^\top$ is a (KDx1) observation vector which describes the short-term and long term f_0 characteristics associated with the n-th syllable, and K the total number of long-term temporal domains being modelled.

A HMM $\boldsymbol{\lambda}_\mathbf{q}$ is estimated for each of the linguistic contexts. Each of the context-dependent HMMs is assumed to be a single-state HMM with single normal distribution and diagonal covariance matrix. Then, a context-dependent HMM $\boldsymbol{\lambda}$ is derived based on Maximum-Likelihood Minimum-Description-Length (ML-MDL). The long-term variations are used as additional trajectory constraints to refine the clustering of the models. A conventional context-dependent HMM is used to model syllable durations.

3.2. Parameters Inference

The determination of the sequence of f_0 parameters is similar to that of the *Trajectory Model* with the exception that the frame-based static observation is reformulated into the stylized f_0 contour over the syllable, and the frame-based dynamic observation (partial derivative) is reformulated into the stylized long-term f_0 contours. The sequence of syllable durations is determined with the conventional static method as the sequence of mean durations.

The optimal static observation sequence $\mathbf{c}$ is determined so as to maximize the log-likelihood of the short-term observation sequence $\mathbf{o}$, under the constraint of the long-term trajectories $\mathbf{\Delta^{(k)}c}$.

The optimal observation sequence $\widehat{\mathbf{o}} = [\widehat{\mathbf{o}_1}^\top, \ldots, \widehat{\mathbf{o}_T}^\top]$ is determined so as to maximize the conditional probability of the observation sequence $\mathbf{o}$ given the model $\boldsymbol{\lambda}$.

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}} \max_{\mathbf{q}} \mathrm{p}(\mathbf{o}|\mathbf{q}, \boldsymbol{\lambda})\ \mathrm{p}(\mathbf{q}|\boldsymbol{\lambda}) \quad (2)$$

The determination of the optimal observation sequence $\mathbf{o}$ divides into the following sub-problems:

$$\widehat{\mathbf{q}} = \underset{\mathbf{q}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{q}|\boldsymbol{\lambda}) \quad (3)$$

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda}) \quad (4)$$

Assuming that each syllable is modelled by a single-state HMM, the optimal state sequence simply corresponds to the concatenated sequence of context-dependent models associated with each syllable of the syllable sequence:

$$\widehat{\mathbf{q}} = [\mathbf{q}_1, \ldots, \mathbf{q}_N] \quad (5)$$

where N denotes is the total number of syllables in the syllable sequence.

The maximization of $\mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda})$ with respect to $\mathbf{o}$ is equivalent to the maximization of $\mathrm{p}(\mathbf{c}|\widehat{\mathbf{q}}, \boldsymbol{\lambda})$ with respect to $\mathbf{c}$ under the dynamic constraints $\mathbf{\Delta}^{(k)}\mathbf{c}$:

$$\widehat{\mathbf{o}} = \underset{\mathbf{o}}{\operatorname{argmax}}\ \mathrm{p}(\mathbf{o}|\widehat{\mathbf{q}}, \boldsymbol{\lambda}) \quad \Leftrightarrow \quad \widehat{\mathbf{c}} = \underset{\mathbf{c}}{\operatorname{argmax}}\ \mathrm{p}(\mathcal{F}(\mathbf{c})|\widehat{\mathbf{q}}, \boldsymbol{\lambda}) \quad (6)$$

under the constraint:

$$\mathbf{o} = \mathcal{F}(\mathbf{c}) = \left[\mathbf{c}^\top, \mathbf{\Delta^{(1)}c}^\top, \ldots, \mathbf{\Delta^{(K)}c}^\top\right]^\top \quad (7)$$

A local solution to this problem is determined recursively using a quasi-Newton method. Finally, global variance is used to model global dynamics [7].

4. Evaluation

4.1. Stimuli

The proposed trajectory model was evaluated and compared to the conventional HMM-based model in a subjective evaluation in speech synthesis. Four models were compared: 1) the conventional HMM-based model (HTS), and trajectory models using different long-term temporal domains: 2) syllable + 1-order syllable-context (1ORDER), 3) syllable + minor prosodic group (AG), and 4) syllable + major prosodic group (PG). Evaluation was conducted using the HMM-based speech synthesis system [1]. Models were trained on 5 hours (1888 utterances) of a French single-speaker story-telling speech database using conventional linguistic contexts. 8 sentences randomly extracted from the fairy-tale "*Le Petit Poucet*" ("*Little Tom Thumb*") were used for the comparison. For each of the trajectory models, the inferred sequence of stylized f_0 parameters was converted into a sequence of f_0 variations with respect to the inferred syllable durations and the voice/unvoiced sequence as inferred from the conventional HMM-based f_0 model. Finally, speech utterances were synthesized by the speech synthesizer. Each sentence was synthesized with the different models.

4.2. Procedure

20 native French speakers (including 13 expert and 7 naïve listeners) participated in the evaluation. The experiment consisted in a subjective comparison of the different speech prosody models. A comparison category rating test was used to compare the *naturalness* of the synthesized speech utterances. The evaluation was conducted according to a *crowd-sourcing* technique using social networks. Pairs of synthesized speech utterances were randomly presented to the participants. They were asked to attribute a preference score according to the *naturalness* of the speech utterances being compared on the comparison mean opinion score (CMOS) scale.

5. Results

Overall CMOS and preference score (PS) are presented in figure 3. The 1-order trajectory model significantly outperforms all of

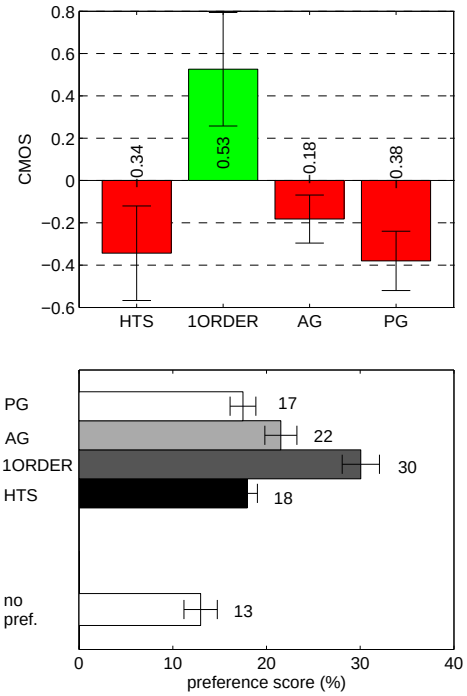


Figure 3: CMOS and PS. Mean and 95% confidence intervals

the other prosodic models whatever is the preference measure. In particular, the 1-order trajectory model is overally significantly preferred to the other prosodic models (CMOS=+0.53, PS=30%), and is individually significantly preferred to each of the other prosodic models (MOS=+0.54,+0.51,+0.54 and PS=52.1%,56.3%,55.1% compared with HTS, AG, and PG models respectively). The AG trajectory model is preferred to the HTS model but not significantly (overall: CMOS=-0.18, PS=22%; pair: CMOS=+0.15, PS=46%); and significantly preferred to the PG trajectory model. Finally, the HTS model is preferred to the PG trajectory model, but not significantly (overall: CMOS=-0.34, PS=18%; pair: CMOS=+0.10, PS=28.7%). In particular, trajectory models decrease in preference when increasing the temporal domain of the trajectory constraint ($\mathrm{CMOS}_{1-\mathrm{order}}$=+0.53,$\mathrm{PS}_{1-\mathrm{order}}$=30%; $\mathrm{CMOS}_{\mathrm{AG}}$=-0.18, $\mathrm{PS}_{\mathrm{AG}}$=22%; $\mathrm{CMOS}_{\mathrm{PG}}$=-0.38, $\mathrm{PS}_{\mathrm{PG}}$=17%).

A comparison of the preference scores depending on the expertise of the participant reveals a significant difference in the perception of speech prosody between naïve and expert listeners : naïve listeners have clearly marked preferences, but with more variability, while expert listeners have less marked preferences, but with less variability (table 1).

CMOS	naive			expert		
	score		rank	score		rank
HTS	-0.77	($\pm$ 0.44)	4	-0.20	($\pm$ 0.27)	2
1-order	**+0.88**	($\pm$ 0.43)	**1**	**+0.41**	($\pm$ 0.26)	**1**
AG	-0.10	($\pm$ 0.50)	2	-0.21	($\pm$ 0.28)	3
PG	-0.20	($\pm$ 0.44)	3	-0.52	($\pm$ 0.24)	4

Table 1: CMOS depending on the expertise of the participant. Mean score and 95% confidence interval.

6. Discussion

A study case of synthesized f_0 variations with respect to the speech prosody model is provided in figure 4 with prior state duration alignment. Speech prosody differences mostly concern f_0 variations, and no significant differences between state-based and syllable-based modelling.

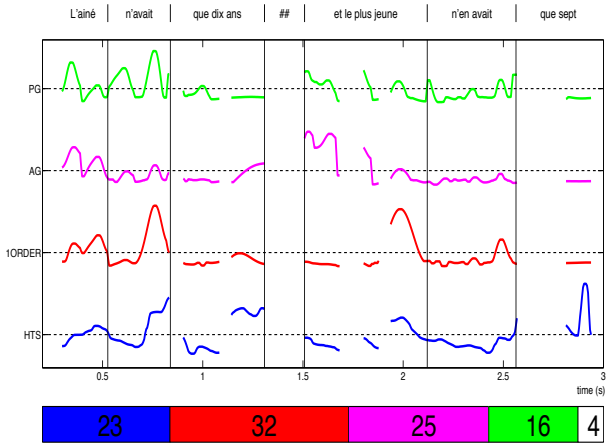


Figure 4: Comparison of synthesized f_0, with PSs.

The 1-order trajectory model clearly succeeds to model the local variations and dynamic of speech prosody. The synthesized f_0 variations presents an expanded dynamics while less micro-prosodic details than those synthesized by the HTS model. Thus, naïve listeners may focus on global variations only, when expert listeners may pay a closer attention to finer prosodic details. The AG trajectory model appears to model middle-term prosodic variations such as initial f_0 reset and local f_0 declination, compared with the 1-order trajectory model and the HTS model. However, dynamics is less expended, and prosodic phrasing is more flat.

A comparison of the different trajectory models reveals that differences in speech prosody concern local (syllable contours and dynamics) and global f_0 variations. However, it is observed that the increase of the trajectory domain results into noisy local f_0 variations, and partially (AG) or totally (PG) inadequate global f_0 contours. In particular, the PG trajectory model failed in modelling global f_0 declination. The degradation is probably due to the increase in the dimensionality of the optimization problem when accounting for long-term trajectory constraints. In the absence of an explicit formulation of the gradient, the optimization method obviously failed to account for the long-term dependencies. Not surprisingly, this results both into local and global degradation in the synthesized f_0 variations.

7. Conclusion

In this paper, a trajectory model based on the stylization and the joint modelling of f_0 variations over various temporal domains was proposed. In the proposed approach, f_0 variations are stylized with a Discrete Cosine Transform, and modelled simultaneously over various temporal domains which cover short-term and long-term variations. During the training, a context-dependent model is estimated according to the joint stylized f_0 contours over the syllable and a set of long-term temporal domains. During the synthesis, f_0 variations are inferred using the long-term variations as trajectory constraints. The evaluation consisted in a subjective comparison of different speech prosody models in speech synthesis.

The 1-order trajectory model was proved to be significantly preferred to the conventional model, and to the other trajectory models. Each of the trajectory models succeeds in modelling f_0 contours that are consistent with the considered temporal domains. However, the ability of the trajectory model to account for long-term variations decreases when the temporal domain increases, due to the increase in complexity of the optimization process. In further studies, the relationship between static and dynamic trajectories will be explicitly formulated, and different combinations of trajectory constraints will be evaluated. Finally, the formulation of the trajectory model will be extend to the modelling of the local speech rate variations.

8. References

[1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] H. Fujisaki, *The Production of Speech*. Springer, New York, 1983, ch. Dynamic characteristics of voice fundamental frequency in speech and singing, pp. 39–55.

[3] J. Van Santen and B. Moebius, *Intonation Analysis, Modelling and Technology*. Kluwer Academic, Netherlands, 1999, ch. A quantitative model of f0 generation and alignment, pp. 269–288.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *European Conference on Speech Communication and Technology*, Budapest, Hungary, 1999, pp. 2347–2350.

[5] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *International Conference on Speech and Language Processing*, Jeju Island, Korea, 2004, pp. 1397–1400.

[6] K. Tokuda, H. Zen, and T. Kitamura, "Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features," in *European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003, pp. 865–868.

[7] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[8] J. Latorre and M. Akamine, "Multilevel parametric-base F0 model for speech synthesis," in *Interspeech*, Brisbane, Australia, 2008, pp. 2274–2277.

[9] Y. Qian, Z. Wu, and F. K. Soong, "Improved prosody generation by maximizing joint likelihood of state and longer units," in *International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009, pp. 3781–3784.

[10] B. Gao, Y. Qian, Z. Wu, and F. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood," in *Interspeech*, Brisbane, Australia, 2008, pp. 2266–2269.

[11] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and Synthesising F0 contours with the Discrete Cosine Transform," in *International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, U.S.A, 2008, pp. 3973–3976.