EFFICIENT GRADIENT F0 TREE MODEL FOR PROSODY MODELING AND UNIT-SELECTION, APPLIED FOR THE EMBEDDED US ENGLISH CONCATENATIVE TTS

Slava Shechtman [1], *Ryuki Tachibana* [2]

[1] Haifa Research Lab., IBM Research, Haifa, Israel
[2] Tokyo Research Lab., IBM Research, Kanagawa-ken, Japan
slava@il.ibm.com, ryuki@jp.ibm.com,

ABSTRACT

Modeling of pitch dynamics in addition to absolute pitch modeling is highly desirable for robust pitch curve prediction and unit selection in concatenative TTS systems. Transition prosody models have been reported to improve consistency and naturalness for pitch-accent and tonal languages, like Japanese and Mandarin. In the current work we revise a Gradient F0 tree model, originally developed for Japanese, and adjust it for American English. The resultant model requires few computational resources at a runtime that makes it highly suitable for embedded TTS applications. We report encouraging results of applying it for an embedded concatenative TTS system for American English.

Index Terms: speech synthesis, unit selection, prosody modeling, F0 modeling, embedded TTS

1. INTRODUCTION

Naturalness has become a main merit of quality of modern state-of-the-art concatenative TTS systems, which generally provide highly intelligible output, while a prosodic aspect of speech synthesis is considered to be essential for natural sounding TTS systems. Prosody is a combination of a number of factors such as fundamental frequency (pitch), duration, energy and pauses. Here we only consider pitch, which is recognized as the most prominent factor for the perception of prosody.

Producing natural F0 curve while preserving high quality of synthesized speech has been always considered a challengeable task for concatenative TTS systems, because judging a specific realization of pitch curve correct for non-tonal languages is not always possible and highly subjective. While pitch accented languages, s. a. Japanese, has a considerable portion of lexical prosodic pattern realizations, which can be unambiguously judged incorrect, in stress-accented languages (or so called intonation languages), s. a. English and many other European languages, there are very few or no such patterns. Those languages rather convey post-lexical pitch accentuation (e.g. high rising intonation in non-wh-questions), based on semantics and high level of speech understanding. They use pitch to convey surprise, irony or enthusiasm, or to express prominence and focus, and their usage of pitch patterns is sometimes oblique and ambiguous.

State-of-the-art concatenative TTS systems generally have rough rule-based or probabilistic models for instant pitch values [1]-[5]. These models are either used both in the unit selection stage and produce the resultant pitch curve for the synthesis [4] (*Target-F0*), or used just in the unit selection, while the actual prosody is derived from the selected units themselves [1][3][5] (*Segment-F0*). In our previous work we reported that incorporation of dynamic F0 observations together with instant pitch observations is beneficial for improving naturalness of large-scale TTS system [7].

In the other work [8], aimed for Japanese prosody improvement, and inspired by a Mandarin prosody model [5], it was proposed to incorporate dynamic and static pitch features into the unit-selection process by separate pitch gradient and absolute pitch GMM modeling and using log-likelihood costs in unit selection process.

In the current work we present the revised and adjusted Gradient F0 model [8] for unit-selection and prosody modeling of American English, combined with the Target-F0 and Segment-F0 combination technique, proposed in [7]. The resultant system showed both subjective and objective quality improvements compared to the baseline system [2], while still keeping its attractiveness for low-cost, low-computation implementations.

The work is organized as follows. First, the conventional CART intonation model is shortly reviewed. Then, the proposed Gradient F0 tree modeling, adjusted for English, will be described. Finally, the results of application of the proposed algorithms to the embedded IBM CTTS system will be presented and discussed.

2. BASELINE F0 MODELING

2.1. Overview

The baseline [1][2] intonation model uses phonetic and semantic features, gathered from the input text, to predict rough pitch curve per syllable. The set of features includes a syllable stress, a word prominence, a part of sentence, syllable and word count, phonetic context and more. A set of 24 features per syllable is extracted over a context window of five syllables, consisting of the current syllable plus the two preceding and two following syllables. From these feature vectors and observations, a decision tree (CART) is built. The observation vector consists of three pitch values

(in log-Hertz), obtained from the beginning of the first syllable's sonorant, the center of the syllable nucleus and the end of its last sonorant. For each leaf of the tree, the average of the F0 distribution of the training data is used as the Target-F0 value at run-time.

The rough Target-F0 curve, predicted from the CART tree, is smoothed and used as a part of the overall cost for the segment-selection dynamic search [1][2], where the basic speech segment for the concatenation is an acoustic HMM state, stored in acoustic contextual decision trees [2]. In order to determine the segment sequence to concatenate, a dynamic programming search is performed over all segments aligned to each leaf of the acoustic decision-trees to minimize a sum of segment costs.

The segment cost consists of a target cost and a concatenation cost. The target cost is the weighted summation of an F0 cost, a duration cost, and an energy cost. These costs are penalties for the differences in the prosodic parameter values of the segments compared to the target prosodic parameter values. For example, the target F0 cost for a speech segment is a penalty for the difference between the Target-F0 value and the Segment-F0 value. The target cost is added up to the concatenation cost, which is the weighted summation of a spectral continuity cost and a F0 transition cost. These costs are the penalties for spectral and pitch discontinuities at segment concatenation points.

2.2. Output F0 curve

There are two basic options for producing output F0 curve, mentioned beforehand, the *Target-F0* and the *Segment-F0*. The first option results in rough and over-smoothed pitch contour, while for the second option the quality of output intonation is heavily dependent on the database size and the correspondence between the database and the synthesized text domains and may result in inconsistent to context output F0 curve.

In [7] we proposed an alternative technique for combination between the Target-F0 and the Segment-F0 curves to keep small intra-syllable pitch fluctuations combined with the rough Target-F0 modeling. This technique proposes improved pitch naturalness compared to the Target-F0 output curve and improved pitch curve consistency compared to the Segment-F0 pitch curve application.

3. THE GRADIENT F0 MODEL

A complementary model of F0 changes (applied together with the conventional F0 model) have been recently developed and successfully applied for tonal and pitch accented languages such as Japanese [8] and Mandarin [5]. Their usage is justified by the fact that most of intonation patterns are expressed by F0 curve dynamics (e.g. raising intonation, lowering intonation, peak, flat, etc.), rather then by F0 absolute values. Those dynamic patterns are extensively used in stress accented languages, such as English as well. In our previous work [7] we proposed using both dynamic and static pitch observations for CART building, and joint dynamic maximum likelihood (Dynamic ML) target curve modeling at a run-time. Although the proposed system resulted in improved Target-F0 curve, it was unsuitable for TTS embedded applications with reduced computational resources. In the current work we adopted a separated dynamic and static F0 features modeling, [5][8], while replacing a computationally intensive Segment-F0 modification towards a target [8] by the Target-F0 and the Segment-F0 combination technique, reported in [7]. After the review of the Gradient F0 modeling [8], we will describe its improvements and adjustments, applied for American English embedded TTS system.

3.1. The tree build

The Gradient F0 model, originally proposed in [8] for Japanese, makes use of two types of CART trees. In addition to the conventional pitch tree, based on three absolute pitch observations per syllable, the complementary gradient tree is built. It is based on three per syllable pitch slope measurements, approximated by linear regression over a fixed period of time (T_s samples ahead of each observation point).

Both absolute and gradient observations are based on cleaned and smoothed pitch data, extracted from the pitch mark data, available for the voice corpus. At the first stage of the process, the *reliable* pitch marks were identified. A pitch mark is labeled as *reliable* if its logarithmic F0 value is not too far from the value predicted by the quadratic curve approximating the preceding three pitch marks. Only the reliable pitch marks are used for the subsequent processing steps. Then, we fill the missing F0 values for the devocalized regions and the unreliable regions by linearly interpolating the logarithmic F0 values of the neighboring reliable pitch marks. Finally, we smooth the obtained logarithmic F0 contours by convolving a Gaussian function to the contours.

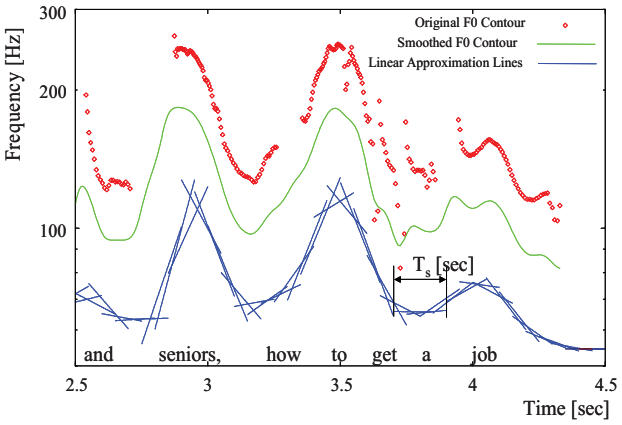

Figure 1: *An example of a smoothed F0 contour and linear approximation. The lines are shifted for better visualization.*

Examples of the smoothing and the linear approximation lines are shown in Fig. 1.

Once the absolute F0 tree and the F0 gradient CARTS are trained, a distribution of each tree leaf data is modeled by a Gaussian Mixture Model (GMM).

3.2. Run time procedure

After the input feature vectors are collected, based on the context information obtained by analyzing the input text, they are used for absolute and gradient F0 trees traversing. Thus, it obtains for later use a set of GMM parameters for each of the three observation point of each syllable. Those models are used to define probabilistic F0-related costs, used in the segment selection procedure. Three new probabilistic costs are used instead of the conventional pitch target and transition costs and summed up with other costs, which are kept the same as in the baseline system [1][2].

Before calculating the F0-related costs for a segment, the closest F0 observation point within the syllable is determined for the segment, and the GMM model, related to that point is used for the cost calculations. The three new costs comprise of 1) the *Absolute F0 cost,* given by $c_{abs,i} = -w_a \log\left(f_{abs,j}\left(\left(p_{start,i} + p_{end,i}\right)/2\right)\right)$, where $f_{abs,j}(p)$ is a GMM probability density, associated with the observation point, closest to the *i*th segment, and $p_{start,i}$ and $p_{end,i}$ are the *i*th segment starting and ending log-pitch values respectively and w_a is a tunable cost weight; the *Gradient F0 cost* and the *Linear Approximation Error cost* [8].

To calculate the Gradient F0 cost, the F0 gradient in the last *Ts*-second period of the current segment sequence is approximated. This is done by combining log-F0 values $\{p_i\}$ of the starting and the ending points of the segments in the interval with their predicted timings $\{t_i\}$ and then calculating the linear regression $l_i(t) = g_i t + s_i$ over those points $\{(t_i, p_i)\}$. Then, the cost is defined as $c_{grad,i} = -w_g \log\left(f_{grad,j}\left(g_i\right)\right)$, where $f_{grad,j}(p)$ is a GMM probability density, associated with the observation point, closest to the *i*-th segment, and w_g is a tunable cost weight.

The *Linear approximation error cost* is given by $c_{app,i} = w_f \sqrt{\frac{1}{N_i}\sum_{k=1}^{N_i}\left(p_k - l_i(t_k)\right)^2}$, and it is used to penalize poor linear approximation.

After the segment selection stage, the original work [8] proposes computationally loaded F0 adjustment stage to modify the Segment-F0 curve (this stage was replaced by the Segment-F0 and Target-F0 combination technique [7], when adopting system for embedded English TTS, so its description is omitted here). Finally, the resultant pitch curve is slightly smoothed and is used for the synthesis of the output speech.

3.3. The Gradient model, revised

3.3.1. The tree build

Several adjustments of the algorithm were made in the tree build and the data collection stage. First, the three pitch observation points were spread over the syllable nucleus, as done in the baseline English system [1][2], rather then evenly spread over the syllable/mora [5][8].

Second, we use 3-dimentional observations for tree building, thus jointly predicting 3 values per syllable (for absolute and gradient values, separately). Each tree leaf was modeled by a single Gaussian model to extract the Target F0 easily to use for final F0 curve generation.

Third, the gradient approximation at a build time was revised. To better fit the gradient approximation at a run-time, we approximate it using only starting and ending pitch values per segment, as it is done at a run-time. However, in the linear regression operation, as described above, a few points are involved (we have roughly 9 segments per syllable), and both long and short segments contribute the same to the slope calculation, so both the build-time data collection and the run-time evaluation should be revised. After the pitch smoothing stage, the smooth pitch curve is sampled at segment joints, so the continuous piecewise linear representation $p(t)$ of pitch curve portion is defined over the T_s time interval. The optimal solution for that regression problem may be formulated by minimizing the continuous squared error function $E(t) = \int_{t_i - T_s}^{t_i} (p(t) - (at + b))^2$ with respect to a and b. Zeroing the derivatives results in:

$$\begin{cases} F_1 \equiv \int_{t_i - T_s}^{t_i} x f(x)dx = \int_{t_i - T_s}^{t_i} (ax^2 + bx)dx = a\frac{t_i^3 - (t_i - T_s)^3}{3} + b\frac{t_i^2 - (t_i - T_s)^2}{2} \\ F_2 \equiv \int_{t_i - T_s}^{t_i} f(x)dx = \int_{t_i - T_s}^{t_i} (ax + b)dx = a\frac{t_i^2 - (t_i - T_s)^2}{2} + bT_s \end{cases} \quad (1)$$

And the optimal solution for the slope approximate gives $a = \left(12F_1 - 6F_2(2t_i - T_s)\right)/T_s^3$, where F_1 and F_2 can be also expressed analytically. This closed analytic solution may be recursively formulated to be applied efficiently at a run-time.

3.3.2. Segment selection

The absolute log-pitch and log-pitch slope data were collected consistently to the build, revised for English as described in section 3.1. The F0 related costs were calculated as described in section 3.2., while the probability distributions were recalculated as follows.

Let t_j, t_{j+1} be observation points, adjacent to the *i*-th segment mid point t_i, $(t_j \le t_i \le t_{j+1})$. Then, knowing the observation data distributions $N(\mu_j, \sigma_j^2), N(\mu_{j+1}, \sigma_{j+1}^2)$, we can evaluate the parameters of the Gaussian distribution at t_i,

assuming that $p_i = \alpha p_{j+1} + (1-\alpha)p_j, \alpha \triangleq \frac{t_i - t_j}{t_{j+1} - t_j}$. In that case, the distribution density at t_i equals to

$$f_i(p) = N(\alpha\mu_{j+1} + (1-\alpha)\mu_j, \alpha^2\sigma_{j+1}^2 + (1-\alpha)^2\sigma_j^2) \qquad (2)$$

The additional revision has to do with more precise log-likelihood calculation. The log-likelihood costs, as formulated in section 3.2., may become negative if not hard-limited or scaled. The scaling factor is unknown at runtime, and limiting makes costs less precise near the peak. In order to resolve this negative costs problem, we replaced probability densities $f_{abs,j}(p_i)$, $f_{grad,j}(g_i)$ by probability approximations $\Pr_{abs,j}(p_i, \Delta p)$, $\Pr_{grad,j}(g_i, \Delta g)$, where Δp, Δg are predefined data vicinities.

3.3.3. Output pitch curve

The Segment-F0 and Target-F0 combination technique, proposed in [7], was used here to combine Segment-F0 with Target-F0. The explicit Target-F0 exists only for single Gaussian modeling of pitch tree leaves (which is the case of the English TTS system settings), and it makes use of the absolute pitch model solely. Both the gradient and the absolute pitch trees are used for the segment selection, so they influence the Segment-F0 curve.

4. EXPERIMENTS AND RESULTS

A low-footprint (10MB) embedded American English TTS system [6] was used to evaluate the proposed algorithm (The Gradient tree modeling, denoted as system C), compared to the baseline CART prosody modeling [1][2] (system A), and the Dynamic ML solution, previously reported in [7] (system B). A set of subjective evaluations has been conducted to assess the perceptual quality, obtained by application of the above. The subjects (7 native English speakers, having no expertise in speech science) listened to 50 sentence pairs each, containing random pairs from the three systems of the above. The subjects were instructed to choose between 5 options: no preference, strong preference to either side or weak preference to either side.

Table 1: *A-B preference tests for the baseline (A), the Dynamic ML (B) and the Gradient F0 (C) systems*

Pref.	No pref.	baseline (A)		Dynamic ML (B)	
		Any pref.	Strong pref.	Any pref.	Strong pref.
%	17.5	24.6	0.9	57.9	19.3
Pref.	No pref.	baseline (A)		Gradient (C)	
		Any pref.	Strong pref.	Any pref.	Strong pref.
%	21.5	21.5	0.8	57	15.7
Pref.	No pref.	Dynamic ML (B)		Gradient (C)	
		Any pref.	Strong pref.	Any pref.	Strong pref.
%	47	25.2	0.9	27.8	4.3

The results are presented at Table 1. It can be seen, that both system (B) and system (C) significantly ($p < 0.05$) outperform the baseline system (A). The differences between the (B) and (C) were found statistically insignificant, with a slight preference towards the system C. So, the computationally efficient separate modeling of F0 static and dynamic features proved to perform at least as good as the joint Dynamic ML solution.

5. SUMMARY

In the current work the F0 Gradient model, first developed for Japanese [8], was revised and adjusted for the embedded English TTS engine [6]. Among main revisions of the system are the optimal piecewise linear gradient estimation (1), linear model interpolation (2) and adopting the Segment-F0 and Target-F0 combination technique, described in [7], rather then computationally costly F0 adjustment [8]. The resultant computationally efficient algorithm, fits well for embedded fixed point platforms and performs at least as good as the joint Dynamic ML solution, previously reported [7].

6. ACKNOWLEDGEMENTS

We would like to thank Mr. Alex Sorin for fruitful discussions, contributed to the algorithm development.

7. REFERENCES

[1] Pitrelli, J. F., Bakis, R., Eide, E. M., Fernandez, R., Hamza, W., Picheny, M. A., "The IBM expressive text-to-speech synthesis system for American English," IEEE Transactions on Audio, Speech & Language Processing, vol. 14, no. 4, pp. 1099–1108, 2006.

[2] Eide, E., *et al.*, "Recent Improvements to the IBM Trainable Speech Synthesis System", *Proc. ICASSP 2003*, Hong Kong, Vol. 1, pp. 708-711..

[3] Raux, A., Black, A. W., "A Unit Selection Approach to F0 Modeling and its Application to Emphasis", *ASRU 2003*, St Thomas, US Virgin Islands.

[4] Hunt, A., Black, A. W., "Unit selection in a concatenative speech synthesis system using a large speech database", in ICASSP '96, Philadelphia, PA, 1996, pp. 373–376.

[5] Ma, X.J., Zhang, W., Zhu, W. B., Shi, Q., Jin, L., "Probability Based Prosody Model For Unit Selection", Proc. ICASSP, Montreal, 2004.

[6] Chazan, D., Hoory, R., Kons, Z., Sagi, A., Shechtman, S., Sorin, A. "Small footprint concatenative text-to-speech synthesis system using complex spectral envelope modeling", Proc. INTERSPEECH-2005, pp 2569-2572.

[7] Shechtman, S., "Maximum-Likelihood Dynamic Intonation Model for Concatenative Text-to-Speech System," in Proc. 6th ISCA Speech Synthesis Workshop, August 2007, pp. 234–239.

[8] Tachibana,, R., Nagano, T., Nishimura, M., "F0 Gradient Model for Acoustic Quality and F0 Consistency of Concatenative TTS," Technical Report of IEICE, 2007-NLC/SP-12, December 2007