# F0 MODELING WITH MULTI-LAYER ADDITIVE MODELING BASED ON A STATISTICAL LEARNING TECHNIQUE

*Shinsuke Sakai*

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
sakai@mit.edu

## ABSTRACT

In this paper, we describe research in fundamental frequency modeling based on a statistical learning technique called *additive models*. A two-layer additive $F_0$ model consists of a long-term, intonational phrase-level component, and a short-term, accentual phrase-level component. It can be learned from the data using a $backfitting$ algorithm, an optimizer of a penalized least-square criterion defined on the model. It estimates two components simultaneously by iteratively applying cubic spline smoothers. To investigate the further flexibility of the model, we incorporated a third additive term that represents a contextual effect on an accentual phrase, and confirmed the improvements in terms of RMS errors. Experimental results on a 7,000 utterance Japanese speech corpus shows an achievement of $F_0$ RMS errors of 28.5 and 29.3 Hz on the training and test data, respectively, with corresponding correlation coefficients of 0.81 and 0.79.

## 1. INTRODUCTION

In recent years, corpus-based concatenative methods for speech synthesis have received increasing attention within the research community, as well as the speech technology industry, because of their ability to generate natural sounding speech output [1]. In general, for synthesized speech to be natural and intelligible, it is crucial to have a proper $F_0$ contour that is compatible with linguistic information such as lexical accent (or stress) and phrasing in the input text. In the corpus-based concatenative speech synthesis setting, target $F_0$ features (e.g., mean frequency, dynamic range) are generated for each synthesis unit. Distance metrics can then be used to compute a cost between the unit target values, and those available in a speech corpus. Overall cost is minimized during search to find the best matching sequence of synthesis units from the corpus.

Regression tree-based approaches are popularly used to predict $F_0$-related measures from a set of linguistic features [2, 3]. A regression tree approach is advantageous in that it is simple to implement, yet powerful. It has a few drawbacks, however. For example, the predicted values do not have a smooth contour, since it essentially represents a piecewise constant function of the input features.

In this work, we propose a simple yet novel multi-layer *additive model* [4, 5] approach to $F_0$ contour prediction, and a method to estimate the component functions through the minimization of a residual sum-of-squares error criterion that include a regularization term. In the following section we explain the additive $F_0$ model by way of a two-layer model example, along with the penalized least-squares criterion and a backfitting algorithm that performs as the minimizer of the criterion. We then describe our new effort to introduce an additional layer to account for a contextual effect on an accentual phrase intonation and the comparative experimental results on a large corpus of Japanese speech.

## 2. ADDITIVE MODEL APPROACH

Similar to previous work that uses parametric forms, e.g. multiple linear regression with indicator variables and second-order linear filters [6, 7], the two-layer $F_0$ model represents the $F_0$ contour, $Y$, as the output of a statistical model that combines a long-range intonational-phrase level component, $g$, and a shorter accentual-phrase level component, $h$:

$$\begin{aligned} Y &= \alpha + g(I, U) + h(A, V) + \epsilon \\ &= \alpha + g_I(U) + h_A(V) + \epsilon, \end{aligned} \tag{1}$$

where $\alpha$ is a constant, $I$ is a discrete-valued input variable that represents a type of intonational phrase, and indexes the relevant function $g_I$. $U$ is a continuous variable representing a time point relative to the starting point of the phrase of type $I$. Similarly, discrete variable $A$ designates a type of accentual phrase, and $V$ represents a time point relative to the starting point of the accentual phrase of type $A$. The

This research was supported in part by an industrial consortium supporting the MIT Oxygen Alliance.

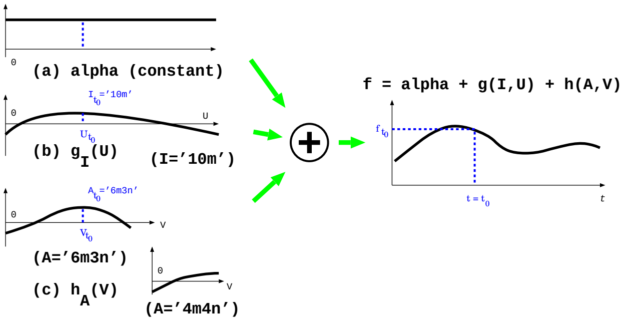


**Fig. 1**. A schematic diagram of a two-layer additive $F_0$ model $f = \alpha + g_I(U) + h_A(V)$. A constant $\alpha$ and component functions $g$ and $h$ are summed up to form the $F_0$ contour $f$.

random error term, $\epsilon$, is zero mean. Figure 1 shows how the three terms form the entire $F_0$ contour function.

A unique characteristic of our approach is that we do not assume any parameterized functional form. Instead, we assume a smoothness defined in terms of curvature, and use an estimation scheme derived from a least-squares error criterion with a regularization term, or roughness penalty [4, 5]. We define the penalized residual sum-of-squares (PRSS) error in the following form:

$$
\begin{aligned}
&PRSS(\alpha, g, h) = RSS(\alpha, g, h) + \lambda_g J(g) + \lambda_h J(h) \\
&= \sum_{n=1}^{N} \{y_n - \alpha - g_{i_n}(u_n) - h_{a_n}(v_n)\}^2 + \\
&\quad \lambda_g \sum_{s \in r(I)} \int g_s''(w)^2 dw + \lambda_h \sum_{t \in r(A)} \int h_t''(x)^2 dx, \quad (2)
\end{aligned}
$$

where $\{(i_n, u_n, a_n, v_n, y_n) | n = 1, ..., N\}$ is a set of training data corresponding to the variables $(I, U, A, V, Y)$, and $\lambda_g$, $\lambda_h$ are fixed smoothing parameters. $r(I)$ and $r(A)$ represents the set of possible values (or *range*) for $I$ and $A$, respectively. The number of elements in a set, for example $r(I)$, will be denoted as $|r(I)|$, hereafter. The first term measures the closeness to the data, while the second and third terms penalize the curvatures in the functions, and smoothing parameters $\lambda_g$ and $\lambda_h$ establish a tradeoff between them. Large values of $\lambda$'s yield smoother curves, while smaller values result in more fluctuation.

It can be shown that the minimizer of (2) is an additive cubic spline model, where $g_I$'s and $h_A$'s are natural cubic splines in the predictor variables $U$ and $V$, with knots, or break points, at each of the unique values of $(i_n, u_n)$ and $(a_n, v_n)$. We can find the solution for (2) with a *backfitting* algorithm [4], a simple iterative procedure depicted in Figure 2.

In the algorithm, we apply a natural cubic-spline smoother, e.g., $\mathcal{S}_i$, to the partial residual, $\{y_{i,l} - \hat{\alpha} - \hat{h}_{a_{i,l}}(v_{i,l})\}_{l=1}^{N_i}$, which is regarded as a function of $u_{i,l}$, to obtain a new estimate $\hat{g}_i$. Partial residual smoothing is done, for $g$'s and $h$'s in turn, using the current estimate of the other component function. The iteration is continued until the estimates $\hat{g}_i$'s and $\hat{h}_a$'s stabilize.

(1) Initialize: $\hat{\alpha} = \frac{1}{N} \sum_{n=1}^{N} y_n, \quad \hat{g}_i \equiv 0, \hat{h}_a \equiv 0$ *for all* $i \in r(I), a \in r(A)$

(2) Cycle: repeat (2g) and (2h) until the functions $\hat{g}_I$ and $\hat{h}_A$ change less than a prespecified threshold.

(2g) Partition the set of training data $\{(i_n, u_n, a_n, v_n, y_n) \mid n = 1, ..., N\}$, into $|r(I)|$ subsets $\{(i, u_{i,l}, a_{i,l}, v_{i,l}, y_{i,l}) \mid l = 1, ..., N_i\}$ $(i \in r(I))$, so that each training point has the same value of $i$ if in the same subset. Note that $\sum_{i \in r(I)} N_i = N$.

For all $i \in r(I)$,

$$\hat{g}_i \leftarrow \mathcal{S}_i[\{y_{i,l} - \hat{\alpha} - \hat{h}_{a_{i,l}}(v_{i,l})\}_{l=1}^{N_i}].$$

(2h) Repartition the training data $\{(i_n, u_n, a_n, v_n, y_n) \mid n = 1, ..., N\}$ into $|r(A)|$ subsets $\{(i_{a,l}, u_{a,l}, a, v_{a,l}, y_{a,l}) \mid l = 1, ..., N_a\}$ $(a \in r(A))$, so that each training point has the same value of $a$ if in the same subset. As before, $\sum_{a \in r(A)} N_a = N$.

For all $a \in r(A)$,

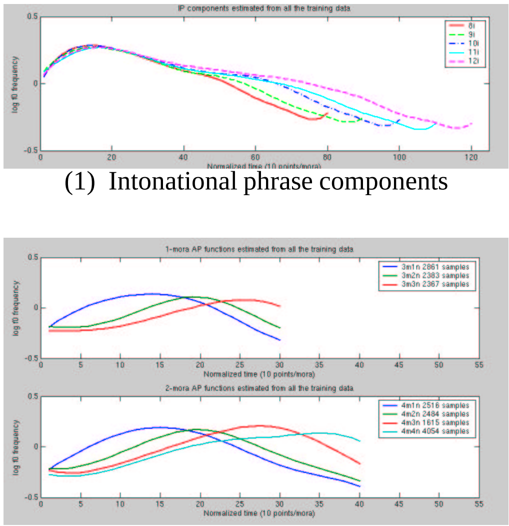
$$\hat{h}_a \leftarrow \mathcal{S}_a[\{y_{a,l} - \hat{\alpha} - \hat{g}_{i_{a,l}}(u_{a,l})\}_{l=1}^{N_a}].$$

**Fig. 2**. A backfitting algorithm for the two-layer additive $F_0$ model.

This backfitting algorithm emerges as a blockwise Gauss-Seidel algorithm for solving a system of linear equations derived from the minimization of the penalized least-square criterion (2) and described in detail in [8].

## 3. ACCOUNTING FOR OTHER FACTORS

We have recently been developing a speech synthesizer for Japanese based on our finite-state transducer-based framework [9], and have created a preliminary version for a weather forecast domain [10]. We have evaluated the use of our $F_0$ modeling technique for Japanese as well. In our initial two-layer formulation, we made a simplifying assumption that an intonational phrase (IP) component of $F_0$ is identified by its mora length. The predictor variable, $I$, represents the number of moras in the IP. An accentual phrase (AP) component is assumed to be identified by the number of moras in it and the position of the nucleus of accent (often called *accent type*). Therefore, the variable $A$ represents a pair $(m, n)$, where $m$ is the number of moras in the accentual phrase and $n$ means that the nucleus is associated with the $n$-th mora. We have had a promising initial



(1) Intonational phrase components

(2) Accentual phrase components

**Fig. 3**. Examples of intonational phrase components and accentual phrase components estimated with the proposed method. (1) Intonational phrase components with the length of 8 through 12 moras. (2) 3- and 4-mora accentual phrase components with all distinct accent nucleus positions.

results from this two-layer additive model [8]. However, as other researchers have pointed out (e.g. [11, 12]), the $F_0$ contour can also be influenced by factors such as word-level context, and segment-level perturbation. To investigate the capability of the additive model framework to incorporate other factors that influences the $F_0$ contour, we have made an attempt to incorporate the effect on the $F_0$ shape of an accentual phrase due to the type of preceding accentual phrase. We introduce a third term in the additive model:

$$Y \quad = \quad \alpha + g(I, U) + h(A, V) + k(B, V) + \epsilon \qquad (3)$$

where, $k(B, V)$ accounts for the effect of the type of preceding accentual phrase. $B$ represents a pair $(m, f)$ where $m$ is the number of moras in the current accentual phrase and $f$ is an indicator which becomes 1 if the preceding accentual phrase has a flat type (i.e. nucleus is on the final mora), and 0 otherwise.

In the backfitting iterations for the three-layer model, we obtain a new estimate of a term, say, $g$, by smoothing the residual of subtracting the current estimates of all the other terms, such as $h$ and $k$, as well as $\alpha$ from the training data, similarly to the two-layer model.

## 4. EXPERIMENTS AND RESULTS

We have implemented the backfitting algorithm for two and three-layer models in Matlab, and estimated component functions $g_i$'s, $h_a$'s, and $k_b$'s in the log frequency domain using a corpus of Japanese utterances read by a female speaker. The corpus comprises 7,282 utterances, which in turn consists of 16,181 intonational phrases (IPs), and 44,717 accentual phrases (APs). A portion of the corpus consisting of 85 intonational phrase was set aside for testing. The number of distinct types of IPs (or distinct mora lengths) was 49, and there were 130 unique AP types. Before the estimation, the original pitch samples were normalized to have the same number of samples per mora. The data instances for which no pitch was extracted for more than half of the mora interval at the beginning or end of all the instances of an AP type were discarded before estimation. The backfitting iteration converged after six loops both for two and three-layer models. As a result, estimates for 46 distinct IPs, 116 types of APs, and 16 functions representing contextual effect on APs were obtained. Figure 3 shows examples of extracted intonational and accentual phrase components.

**Table 1**. Experimental results for two-layer and three-layer additive $F_0$ models

| | RMSE(train) | Corr(train) | RMSE(test) | Corr(test) |
|---|---|---|---|---|
| 2-layer | 28.9 | 0.806 | 29.8 | 0.777 |
| 3-layer | 28.5 | 0.812 | 29.3 | 0.786 |

As an objective evaluation, we measured the goodness of fit in terms of root mean square error (RMSE) and correlation coefficient (Corr), which are often used in the evaluation of $F_0$ modeling [2, 13]. In the two-layer model, RMSE was 28.9 Hz, and the Corr was 0.806 for the training data, and measured on 85 intonational phrases set aside from the training data, RMSE and Corr were 29.8 Hz, and 0.777, respectively. The standard deviation of the corpus $F_0$ itself was 48.2Hz.

As shown in Table 1, the three-layer model improved the overall RMSE for both training and test sets, although Corr metrics showed a rather small improvements. From a significance test in which mean square errors from two models are regarded as sample variances from two unknown normal distribution, we confirmed that the mean square errors of the two-layer and three-layer models are significantly different with $\alpha = 0.05$.

Figure 4 illustrates an example of the $F_0$ contours from two-layer and three-layer additive $F_0$ models plotted with the actual $F_0$ data in the test set. We see from the figure that the elevation of starting $F_0$ values influenced by the flat shape of $F_0$ for the preceding accentual phrase which is not followed by the two-layer model is nicely accounted for by the three-layer model (see near the cursor at time 1.75).

Although it can be difficult to compare performance across different speech corpora and languages, we believe these results are comparable to state-of-the-art results of 33–34 Hz RMSE, and 0.6–0.72 Corr, that have been reported on a female-speaker English radio news corpus [13, 2] with the standard deviation reported as e.g. 53Hz in [13].

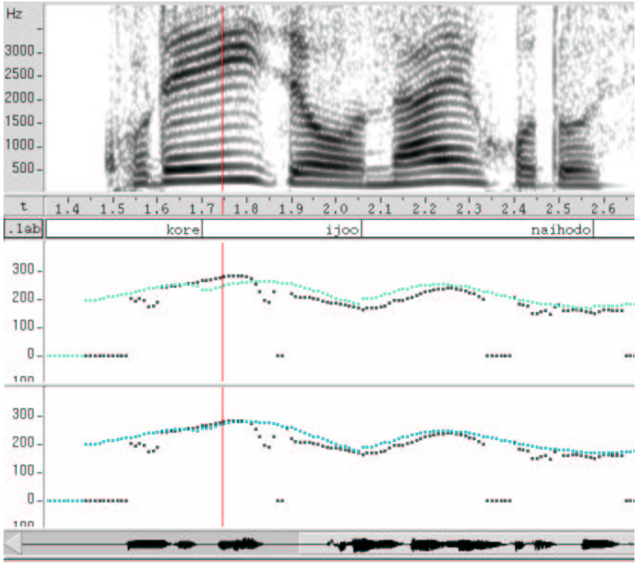


**Fig. 4**. $F_0$ contour from the trained models, displayed with the actual $F_0$ contour of a test data. Output from the two-layer model is shown below the spectrogram, and the $F_0$ from the three-layer model is shown at the bottom. The dark dots are the $F_0$ data in the corpus, and light dots are the $F_0$ contour derived from the additive model.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel multi-layer approach to $F_0$ modeling, and have estimated intonational and accentual phrase components, as well as a component that account for a contextual influence, from a Japanese speech corpus. The fundamental frequency predicted by the model can be used as the reference for deriving a substitution (target) cost for unit selection in a corpus-based speech synthesizer. It may also be used as part of a post-processor to modify the waveform units to have pitch contour closer to the target. We plan to incorporate the $F_0$ measures predicted by the model, as one of the target measures to derive the costs, into our speech synthesis system. We also plan to apply this framework for $F_0$ modeling of English, for more general purpose concatenative speech synthesis.

## 6. ACKNOWLEDGMENT

The authors is grateful to Michael Phillips, Dan Faulkner, Bill Ham, and Yun-Sun Kang at Scansoft for providing the annotated Japanese speech corpus as well as helpful information, and to Jim Glass for the insightful comments.

## 7. REFERENCES

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP '96*, 1996.

[2] X. Sun, "F0 generation for speech synthesis using a multi-tier approach," in *Proc. ICSLP 2002*, Denver, 2002, pp. 2077–2080.

[3] M. Chu, H. Peng, H. Yang, and E. Chang, "Non-uniform units from a very large corpus for concatenative speech synthesizer," in *Proc. ICASSP 2001*, 2001.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.

[5] T. Hastie and R. Tibshirani, *Generalized Additive Models*, Chapman and Hall, 1990.

[6] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan(E)*, vol. 5, no. 4, pp. 233–241, 1984.

[7] M. Abe and H. Sato, "Two-stage F0 control model using syllable based F0 units," in *Proc. ICASSP '92*, San Francisco, 1992, pp. 53–56.

[8] S. Sakai and J. Glass, "Fundamental frequency modeling for corpus-based speech synthesis based on a statistical learning technique," in *Proc. ASRU 2003*, St. Thomas, 2003.

[9] J. Yi and J. Glass, "Information-theoretic criteria for unit selection synthesis," in *Proc. ICSLP 2002*, Denver, 2002, pp. 2617–2620.

[10] M. Nakano, T. Minami, S. Seneff, T. J. Hazen, D. Scott Cyphers, J. Glass, J. Polifroni, and V. Zue, "Mokusei: A telephone-based Japanese conversational system in the weather domain," in *Proc. European Conf. on Speech Communication and Technology*, 2001.

[11] J. Bellegarda, K. Silverman, K. Lenzo, and V. Anderson, "Statistical prosodic modeling: from corpus design to parameter estimation," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 52–66, 2001.

[12] Y.Yamashita, T.Ishida, and K.Shimadera, "A stochastic f0 contour model based on clustering and a probabilistic measure," *IEICE Transactions on Information and Systems*, vol. E86-D, no. 3, pp. 543–549, 2003.

[13] K. E. Dusterhoff, A. W. Black, and P. Taylor, "Using decision trees within the tilt intonation model to predict F0 contours," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, 1999, pp. 1627–1630.