# A Dynamical System Model for Generating Fundamental Frequency for Speech Synthesis

Kenneth N. Ross, *Member, IEEE,* and Mari Ostendorf, *Senior Member, IEEE*

***Abstract*— Higher quality speech synthesis is required for widespread use of text-to-speech (TTS) technology, and prosody is one component of synthesis technology with the greatest need for improvement. This paper describes a new approach to generation of two important cues to prosodic patterns—fundamental frequency $(F_0)$ and energy contours—given symbolic prosodic labels and text. Specifically, the approach represents vectors of $F_0$ and energy with a dynamical system model, which allows automatic estimation of parameters from labeled speech. Parameters at different time scales in the model are structured to capture segment, syllable, phrase and discourse level effects based on linguistic research. $F_0$ generation experiments with the dynamical system model show improved synthetic speech quality over the hybrid target/filter approach.**

***Index Terms*—Dynamical system model, $F_0$ generation, intonation modeling, prosody synthesis.**

## I. Introduction

TEXT-TO-SPEECH (TTS) technology is currently useful only in a limited number of applications because synthetic speech quality is not sufficiently high. Prosody, which includes the phrase and accent structure of speech, is one of the least developed parts of existing systems for converting text into speech, according to Klatt [1], Collier [2] and others. Prosody aids the listener in interpreting an utterance, by grouping words into larger information units and drawing attention to specific words, and Silverman [3] has shown that a domain-specific model of prosody can significantly improve the comprehension of synthesized speech. Since fundamental frequency $(F_0)$ and energy are important acoustic correlates of many prosodic constructs, the goal of this research is to develop an automatically trainable model for generating $F_0$ and energy contours that can be incorporated into existing TTS synthesis systems to improve the naturalness and intelligibility of computer-generated speech.

Linguistic theory supports the practice of modeling prosody by predicting symbolic markers from text (for prosodic phrase boundaries, accents, and the "tones" (or pitch events) that mark them), and then using these markers to generate pitch, energy, and duration patterns. Working within this framework, we describe a model for the $F_0$ and energy generation component of a TTS system. We assume the existence of a text processing module that can provide lexical information (phonemic representations and lexical stress) and symbolic prosodic markers. Other research efforts are concerned with these problems and are relatively successful, e.g. [4], [5] for prediction of prosodic phrases and [6], [7] for prediction of pitch accent location.

For this research, the general approach for developing models of intonation and energy is statistical rather than the more typical rule-based approach. Specifically, we combine a regression tree for $F_0$ range prediction with a stochastic dynamical system model for contour generation, where parameters of both components are automatically trained from a labeled corpus. However, linguistic theory is also used to help define the overall structure and parameter dependencies for the model, because statistical techniques work best when appropriate structures have been chosen and important features are known prior to modeling. For example, the model incorporates four kinds of phenomena that affect at least the $F_0$ contour at the level of: discourse, phrase, syllable and segment. Thus, elements of the model can be related to more traditional rule-based approaches to intonation modeling. Parameter estimation is based on the iterative maximum likelihood algorithm developed by Digalakis *et al.* [8], which is extended here to handle partially unobserved data (unvoiced regions) and parameter tying that varies over different time scales. Experimental results are based on a corpus of FM radio news announcer speech collected at Boston University, which was hand-labeled with prosodic markers. FM radio news speech appears to be a good style for research in prosody synthesis, because the announcers try to sound natural while reading with a well-constrained communicative intent.

Our approach to modeling $F_0$ and energy has a number of advantages over other reported work. The mathematical model itself is theory neutral, while it allows for integration of linguistic knowledge that may build on a particular theory. The use of a statistical model, in particular, is attractive because it can represent the natural variability in prosodic patterns as well as learn prosodic behavior that is not yet understood linguistically. By using a model that can be automatically trained, the parameters can be easily modified to different speaking styles and different languages through retraining on a different set of prosodically labeled speech data. In addition, a statistical model can easily be used in both prosody recognition and synthesis applications.

The body of this paper is organized as follows. Section II reviews other $F_0$ generation work and some results from

Manuscript received June 29, 1995; revised May 14, 1998. This work was supported in part by the National Science Foundation under Grant IRI-8905249 and by a research Grant from NYNEX. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Kahn.

K. N. Ross is with Alphatech, Inc., Burlington, MA 01803 USA (e-mail: kross@alphatech.com).

M. Ostendorf is with the Electrical and Computer Engineering Department, Boston University, Boston, MA 02215 USA (e-mail: mo@bu.edu).

Publisher Item Identifier S 1063-6676(99)02742-X.

research in linguistics that affected our model design. Next, Section III discusses the dynamical system model and its application to $F_0$ and energy generation. Section IV gives details on training the dynamical system model, including some special considerations when using it to model $F_0$ and energy. Section V presents the experimental results, both quantitative and perceptual, and Section VI concludes with a brief summary.

## II. Background

This section reviews background research that motivates our approach to generating $F_0$ contours and provides the basis for many of the parameter dependencies used in the experiments. We begin with a brief discussion of the factors that affect $F_0$ contours and the prosodic theory that our particular implementation is based on. Then we summarize related approaches to $F_0$ modeling, indicating how our work builds on and extends these models. (Few recent studies have looked at the behavior of energy and prosodic structure, perhaps because the relation of energy to loudness is not as well understood as that between $F_0$ and pitch, hence the focus on $F_0$ here.)

### A. Features of the $F_0$ Contour

Effects on the $F_0$ contour can be distinguished according to the size of the linguistic unit they affect, which may be at the discourse, phrase, syllable, or segmental level. We separate these effects into two groups: those that are due to tonal (accent and phrase) structure and operate at the syllable and phrase levels to create an abstract *intonation contour*, and those that modulate the intonation contour to give the observed $\mathbf{F}_0$ *contour* and that operate at the discourse and segment levels.

Realization of the different possible types of accent and boundary markers in English is generally more local, and in most cases can be modeled at the *syllable level*. Both pitch accents on prominent syllables and the edges of phrases are marked by a variety of tones, e.g., relatively higher or lower pitch. The realization of the tones depend on the specific tone type and on factors affecting the timing of the tones relative to the segments. The syllable's structure (number of sonorant segments) affects timing, though use of syllable structure can be difficult because of ambiguity of syllable boundaries in some cases (e.g., the /n/ in "dinner"). Silverman and Pierrehumbert [9] find that pitch accent peak alignment is affected by a number of factors including the distance in time to the next tone (either a pitch accent or a boundary tone), though not affected by the prosodic labels of previous syllables.

At the *phrase level*, theories of intonation must account for the observation that, on average, $F_0$ later in the phrase is lower. Traditionally, this phenomenon was attributed to a declining baseline (e.g., [10]), although some have attributed it to a prominence lending fall in a common phrase-final tune [11]. Liberman and Pierrehumbert [12] returned to the original observation that *peaks* can decline too, and so separated tone-specific effects from phrase-internal pitch range and baseline manipulations of $F_0$ peak downstep and final lowering proper. Beckman and Pierrehumbert [13] further explore these questions and suggest that the appropriate unit for representing pitch range is the intermediate intonational phrase, providing a theory that we will roughly follow here. Linguistic theory also holds that the last accent in a phrase, the nuclear accent, tends to be perceived as more prominent than earlier (prenuclear) accents. Other phrase-level effects include accent timing, which differs in phrase-final position, and the relatively high incidence of glottalization (and therefore $F_0$ tracking errors) in phrase-final position.

By the *discourse level*, we mean prosodic effects that span several phrases, such as the correlates of discourse topic structure. (See [14] for a summary.) Assuming that discourse-level effects on accent placement and choice of tones is modeled in other TTS components, the main impact of discourse structure on realization of $F_0$ is pitch range variation. Phrasal peaks lower gradually within a coherent discourse unit and are reset at discourse segment boundaries, with extra final lowering at the ends of some units (e.g., [15], [16]). Disruption to this pitch range downtrend may cue discourse structures outside of the narrative flow, such as parenthetical phrases [17].

Silverman's thesis [18] includes an extensive discussion of *segment level* effects on the $F_0$ contour (sometimes referred to as microprosody), where he argues that segmental effects on $F_0$ are important for natural sounding intonation. His study found that the effect of consonants on $F_0$ are often large and can be distinguished by listeners from sentence intonation, and that listeners adjust for vowel-related $F_0$ variation in perception of intonational prominence. A study of segmental effects on timing and $F_0$ in pitch accents by van Santen and Hirschberg [19] finds strong effects of the consonant class on the following vowel. They find similar but smaller effects for consonant perturbation on unaccented syllables. Additionally, for accented syllables only, they find the vowel height strongly affects the $F_0$ contours (high vowels produced higher pitch values than low vowels), as shown in earlier work as well.

The model for $F_0$ generation developed in this paper builds on much of the research summarized in this section, taking the ToBI (Tones and Break Indices) prosodic labeling system [20]–[22], as the symbolic representation of the intonation contour. The model represents four levels of prosodic effects through parameter dependence specifications, as described in Section III-B, but the general shape and size of these effects are learned automatically from labeled training data. Timing effects are captured in part through automatic training of multiple models that differ in prosodic context, and in part through specification of a deterministic time warping function within the syllable to capture durational effects on timing.

### B. Models for Generating $F_0$ Contours

Many researchers have developed models for $F_0$ contour generation in a number of languages. There are two important issues that distinguish the various models: whether or not they represent additive effects (illustrated in Fig. 1), and whether the low-level units are generated parametrically or

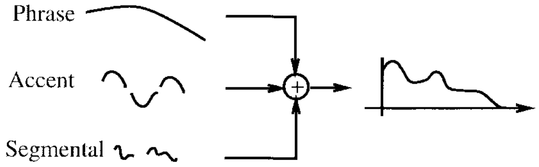
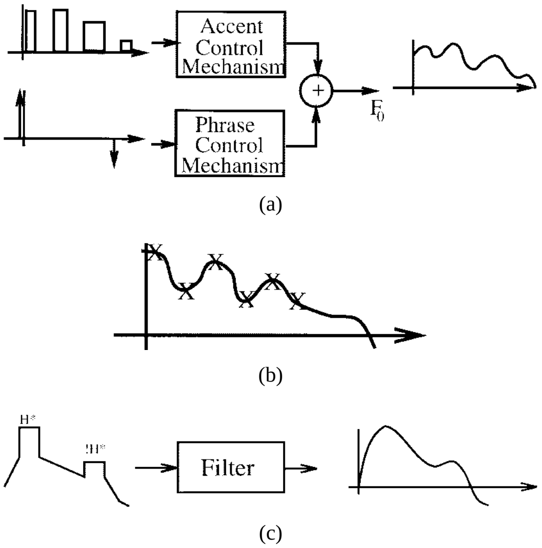


Fig. 1. Possible additive components of a model for generating $F_0$ contours.



(a)

(b)

(c)

Fig. 2. Diagram showing the different types of interpolation and/or filtering models for generating $F_0$: (a) the command/filter model, (b) the target/interpolation model, and (c) the hybrid model.

by concatenating templates. Those models that do represent segmental effects (e.g., [18]) generally use an additive model, and we shall do the same. Parametric models facilitate use of linguistic knowledge and have often been rule-based (e.g., [2], [10], [23]), while template models are typically trained automatically (e.g., [24]). Both approaches can in principle be trained automatically, but the parametric approach can better take advantage of linguistic knowledge through imposed constraints on parameter dependencies. Many parametric approaches to $F_0$ modeling generally fall under one of three categories (illustrated in Fig. 2): 1) command/filter models, 2) target/interpolation models, and 3) hybrid target/filter models that combine features of the other two models. Examples of each are described below, since our model builds on the hybrid approach.

Fujisaki and colleagues [25], [26] have developed a command/filter model to generate $F_0$ contours for Japanese by smoothing a command signal with a linear filter. In Fujisaki's model, linear filters are driven by discrete commands that are generated from the syntactic and lexical information contained in the sentence. The model has two main components [Fig. 2(a)]: word accents are represented by square pulses of varying amplitudes and durations applied to a second-order filter, and higher phrase level effects are represented by impulses applied to a different second-order filter. Final lowering can be incorporated by negative-going impulses at the end of the phrase, although in many languages it may not be consistent with the physiological motivation of the Fujisaki model. This model has been extended to several languages.

Pierrehumbert [23] proposed an $F_0$ model that operates by assigning a series of target values that are then connected together by a parabolic transition function [Fig. 2(b), where "X" indicates targets], improving on earlier work that used linear interpolation to connect targets. More recent target-interpolation models [27], [28] use a general spline function interpolation, and the targets are learned automatically and not tied to a phonological model. In particular, the model proposed by Aubergé [28] incorporates a multilevel representation, similar to that proposed here, but including more phrase levels.

Anderson *et al.* [29] have developed a rule-based parametric $F_0$ model that takes a hybrid approach to synthesizing English $F_0$ contours. As in the earlier Pierrehumbert model, the targets are specified sequentially according to rules that draw on phonological theory [12], [23], [30]. Like the Fujisaki model, a filter is used to smooth target values. However, rather than a sequence of square pulses, the input to the filter is a sequence of short constant target levels connected linearly at the endpoints [Fig. 2(c), where the targets are the level regions], allowing for faster changes in shorter duration contours. Silverman [18] expanded on this model by incorporating additive segmental effects and representing a set of rules for phenomena that span multiple phrases. The resulting full model captures the different linguistic phenomena that we are interested in here and provides the basis for our particular multilevel approach. However, it has the disadvantage that it requires a high cost in the manual effort of rule development. Thus, our goal was to develop a model that could be automatically trained, but retained as much of the framework of the hybrid target/filter models as possible.

## III. A Dynamical System Model of $F_0$ and Energy

The dynamical system model is a general method of modeling a linear time-varying process, and Section III-A begins by reviewing the basic mathematical framework, comparing the dynamical system model to the hybrid target/filter model to make it clear how the new model applies to the problem of $F_0$ and energy generation. Section III-B supplies details of how the model structure is constrained to conform to theoretical models of intonation, and Section III-C covers issues of timing control. The dynamical system model represents a normalized contour, where discourse-level range effects are removed by scaling the $F_0$ and energy with respect to the peak for the intermediate phrase. In the final output, this contour is then scaled back to Hz by a range term predicted with regression trees, as discussed in Section III-D. Most of the discussion in this section is generally applicable, but some of the parameter dependencies are determined by the specific phonological theory adopted here and/or by training limitations associated with the corpus used in these experiments.

### A. General Mathematical Framework

The dynamical system model mathematically represents a sequence of observations $\mathbf{Y} = [y_0, y_1, \cdots, y_N]$ in terms of a sequence of unobserved states $\mathbf{X} = [x_0, x_1, \cdots, x_N]$ where both $x_k$ and $y_k$ can be vectors. In this application, the observation $y_k$ is a two-dimensional (2-D) vector containing values for $F_0$ and energy, and the state $x_k$ is also a 2-D vector. (An advantage of having a joint vector model, rather than separate scalar models for $F_0$ and energy, is that correlation between the two features is represented.) The general form for the discrete-time state-space model is as follows:

$$x_{k+1} = F_j x_k + u_j + w_k \tag{1}$$

$$y_k = H_j x_k + b_j + v_k \tag{2}$$

where, in (1), $u_j$ is an additive, deterministic input to the state equation, and in (2), $b_j$ is an additive, deterministic input to the observation equation. The terms $w_k$ and $v_k$ are zero-mean, uncorrelated Gaussian noises with covariances $Q_j$ and $R_j$, respectively. The initial state is assumed to be $x_0$, where $x_0$ is Gaussian (with mean $m_{x0}$ and covariance $\Sigma_{x0}$) and independent of the observation and state noises. The subscript $j$ selects the set of model parameters appropriate for a particular $y_k$, as will be described further below.

Ignoring the noise terms, the state equation is simply the equation for a first-order filter, where the input is $u_j$ (which can be thought of as a sequence of target values), and the observation equation provides for modeling multiplicative and additive effects (e.g. pitch range and segmental perturbations, respectively). Thus, the model is similar to the hybrid target-filter models [18], [29], except that there is no linear interpolation between the constant targets and, more importantly, it includes the random terms $w_k$ and $v_k$.[1] The Gaussian vectors $w_k$ and $v_k$ can be thought of as representing modeling error, i.e., the uncertainties inherent in the mathematical model due to our limited linguistic understanding of intonation and due to approximations made for computational tractability. The Gaussian vectors $v_k$ can also capture observation noise associated with pitch tracking errors. In addition, by making the model explicitly stochastic, we can benefit from existing maximum likelihood training algorithms, and can apply the model to automatic prosody tone recognition tasks. Another potential advantage that the dynamical system model has over other hybrid target/filter approaches is that it provides for a joint model of $F_0$ and energy.

Since the target values (at least) change as a function of time, the model parameters (as indexed by $j$) change and a mapping $j = \rho(k)$ is needed to specify which parameters to use at any given time. In the case of $F_0$ modeling, the $\rho(k)$ would be determined by syllable, phrase, and segment level dependencies. The mappings between $k$ and sets of model parameters need not be the same for each of the parameters $H_j$, $b_j$, $F_j$, and $u_j$, but since nonuniform mappings require only a simple extension of the derivations here in combination with a significantly more complex notation, the simpler notation is used here and the means for extending to the more general case is discussed in Section IV. Further details on the parameter dependencies explored here are given in Section III-B, and the sequential timing control provided by $\rho(k)$ is explained further in Section III-C.

In speech processing, the dynamical system model was originally proposed for recognition of phonemes [8] and it has also been used for prosodic tone recognition [40]. However, it can also be used in the generation mode by setting the model's random noise components to zero, selecting the appropriate sets of model parameters, and using the resulting predicted means of the stochastic process:

$$x_{k+1} = F_j x_k + u_j \tag{3}$$

$$y_k = H_j x_k + b_j. \tag{4}$$

The subscript $j$ indicates the set of model parameters appropriate for $y_k$, which are are derived from the control function, $j = \rho(k)$, taking as input the predicted prosodic label sequence, syllable structure and lexical stress, and predicted segment duration. The model generates $F_0$ and energy values for all times $k$, and those values for $F_0$ in unvoiced regions are ignored. Lastly, when the model is trained on data normalized for pitch and energy range, as proposed here, the resulting values must be scaled according to the predicted range.

### *B. Model Parameter Dependencies*

An important part of development of the dynamical system model of $F_0$ and energy is the proper selection of model parameter dependencies. Ideally, all of the model parameters would depend on all possible factors and the estimation algorithm would find the correct way to incorporate these factors into the model. But, to overcome limitations of training data and problems with local optima, the parameters are constrained based on current linguistic theory, as reviewed in Section II-A.

Parameter dependencies are represented within the model by the terms $\beta_k$, $\alpha_k$, $\pi_k$, and $\gamma_k$, where the parameters are as follows:

$\beta_k$ position within the intermediate phrase of the syllable containing frame $k$;

$\alpha_k$ prosodic markers on the syllable containing frame $k$;

$\pi_k$ prosodic markers on syllables in the phrase after the one containing frame $k$;

$\gamma_k$ the phoneme class of frame $k$.

The dynamical system equations with the parameter dependencies explicitly marked are as follows:

$$x_{k+1} = F(\alpha_k)\, x_k + u(\alpha_k, \pi_k) + w_k \tag{5}$$

$$y_k = H(\beta_k, \gamma_k)\, x_k + b^1(\beta_k) + b^2(\alpha_k, \gamma_k) + v_k. \tag{6}$$

The noise components of the model, $w_k$ and $v_k$, are zero-mean, uncorrelated Gaussian noises with covariances of $Q$ and $R(\beta_k)$, respectively. The scaling matrix $H(\beta_k, \gamma_k)$ in the observation equation is constrained to be diagonal to reduce the number of free parameters. Parameter sharing is used for the noise covariances to allow the data to be used more efficiently, across the syllable regions and prosodic labels for $Q$ and across the syllable regions and phoneme classes for $R(\beta_k)$.

The parameter dependencies in the dynamical system model are essentially divided in two groups. The state equation (5) models the abstract intonation contour, which depends on tone labels $\alpha_k$ and their prosodic context $\pi_k$. The state equation can be thought of as a representation of the hybrid target/filter

[1] Any differences due to filter order can be accommodated by increasing the vector dimensionality to represent a higher-order process.

model, with $u$ specifying the targets and $F$ specifying the smoothing filter. The $F_0$ contour is generated from the intonation contour according to the observation equation (6), which includes additive segmental effects and pitch range/baseline manipulations in the $b$ and $H$ terms. Relative prominence, as distinct from tone type, is typically implemented with peak scaling rules, and can be controlled in the dynamical system model by either scaling the targets $u$ or the smoothed contour via $H$ (prior to final range scaling).

The additive term in the observation equation is split into $b^1(\beta_k)$ and $b^2(\alpha_k, \gamma_k)$, to simplify parameter estimation by separating phrase-level effects from those at the phoneme and syllable level and to reduce the number of free parameters. As described next, the parameters $\beta_k$ (phrase position), $\alpha_k$ (tone label), and $\gamma_k$ (phoneme class) have 3, 30, and 5 categories, respectively. Thus, a combined term would have $3 \times 30 \times 5 = 450$ versus $3 + 30 \times 5 = 153$ parameters for separate $b^1(\beta_k)$ and $b^2(\alpha_k, \gamma_k)$, respectively. The division is also motivated because it eliminates some complex dependencies that are not consistent with linguistic theory and empirical observations, i.e., the size of the additive segmental effects is affected by both the prosodic label $\alpha_k$ on the syllable and the phoneme class $\gamma_k$ but does not depend on the position of the syllable within the phrase. Thus, segmental effects can be modeled by the additive term $b^2(\alpha_k, \gamma_k)$, while the term $b^1(\beta_k)$ captures phrase-final lowering.

Below, we describe details and motivation for the specific choice of parameter dependencies used in our experiments, including both linguistic motivation and experimental assessment. In many of cases, the number of categories was restricted because of training limitations (see Section V-A).

*1) Phoneme Level:* Five *phoneme categories,* $\gamma_k$, are used to capture segmental effects: vowel, sonorant, voiceless obstruent, voiced obstruent, and strong fricative. These classes were selected primarily to group phonemes with similar segmental effects on the $F_0$ contour [18], with strong fricatives (such as /s/, /z/) separated from the other obstruent consonants (such as /t/) to improve energy modeling for these phonemes. The additive term $b^2(\alpha_k, \gamma_k)$ incorporates segmental effects while accounting for interaction with prosodic labels [18] through the additional dependence on $\alpha_k$. The term $H$ depends on $\gamma_k$ to provide a mechanism for capturing multiplicative segmental effects, though ideally (given more data) $\gamma_k$ should distinguish between high and low vowels to capture differences in vowel intrinsic pitch [18]. Segmental effects on timing could be modeled by incorporating phoneme level effects into the additive term $u$ in the state equation, but experimental evaluation found this hurt performance, probably because of the increased number of parameters. Segmental effects on timing may be more efficiently captured by the distribution mapping function described in Section III-C. Initial values for the parameters for the EM algorithm for $b$ and $\gamma$ were obtained from Silverman's model [18].

*2) Syllable Level:* Parameter dependencies based upon the *prosodic markers on syllables,* $\alpha_k$, are derived from the syllable's pitch accents and phrase tones. To model the data available (see Section V-A), we use five categories of syllable labels, two unaccented (lexically "unmarked" and "stressed") and three accented ("high" H*, "downstepped high" !H*, and "low" L*), and six categories of phrase tones ("none," "L–L%," "L–H%," "L–," "H–," and "!H–").[2, 3] This means that $\alpha_k$, the prosodic markers on syllables, can be one of 30 categories. The primary role of $\alpha_k$ is in determining the intonation contour, and it therefore appears in $F(\alpha_k)$ and $u(\alpha_k, \pi_k)$ in the state equation, as well as in $b^2(\alpha_k, \gamma_k)$ as explained above. The targets $u(\alpha_k, \pi_k)$ also depend on the *prosodic markers on adjacent syllables,* $\pi_k$, since prosodic context influences accent timing. Since the presence of a boundary tone on the same syllable is captured in the definition of $\alpha_k$, the main role of $\pi_k$ is to capture the effect of accent specifications or boundary tones on neighboring syllables. For the results reported here, $\pi_k$ takes on three possible values depending on whether the current syllable has a nuclear accent (last pitch accent in the phrase) or is postnuclear, is prenuclear and followed by an unaccented syllable, or is prenuclear and followed by an accented syllable. Other classifications of prosody on surrounding syllables tended to improve the RMS errors for the training data but increased test data errors. A larger training set would make it possible to experiment with more syllable labels (both accented and unaccented) and more detailed context specification.

*3) Phrase Level:* The *position of the word in the phrase,* $\beta_k$, is incorporated into the observation equation through the scaling term $H(\beta_k, \gamma_k)$, the noise covariance term $R(\beta_k)$, and the additive term $b^1(\beta_k)$, using three regions for the variable $\beta_k$: first word, middle words, and last word. The position in the phrase is used in the scaling term $H(\beta_k, \gamma_k)$ to model effects such as decreasing peak height during the course of a phrase. Incorporating phrase dependence in $u(\alpha_k, \pi_k, \beta_k)$ did not significantly improve performance, and so was abandoned in the interest of decreasing the number of free parameters. The observation noise covariance, $R(\beta_k)$, is dependent on the position in the phrase because pitch tracking errors occur more frequently at the ends of phrases, where especially low pitch is realized as creak. The position in the phrase is used in the additive term $b^1(\beta_k)$ to model potential effects arising from $F_0$ final lowering. (The term could also represent baseline declination if it was specified as a linear slope.)

### *C. Parameter Timing Control: Mapping Distributions to Observation Times*

One difficulty with intonation modeling is that factors operating at different time scales influence the $F_0$ contour. The observation feature vectors ($F_0$ and energy) occur at the frame level (e.g., every 10 ms), but the parameters of the model change in association with changes in phonetic segment category, syllable tone type, position of the syllable in the phrase, and with pitch range changes at phrase boundaries. Here, the pitch range changes are handled by a separate

[2] ToBI tone labels are used here, but in fact the labels represent broader classes as described in Section V-A.

[3] Theoretically, phrase accents can spread over many syllables and are not associated with a single syllable. For practical reasons, we chose to make it a syllable-level label as well, which is reasonable for the radio style (used here) that has short phrases. For longer phrases, one can associate this label (or context-dependent versions of this label) with all post-nuclear syllables in the phrase.

scaling step after the dynamical system model generates a normalized contour. The segment-related parameter changes are synchronized with phone boundaries, based on automatic alignments in training and independently predicted durations in synthesis.

Effects of tone type $(\alpha_k)$ and position in the phrase $(\beta_k)$ operate at the syllable level but have time-varying patterns within the syllable. This variation is accounted for with a finite number of constant targets through: 1) the definition of "regions" of a syllable, where a region has constant parameters ($F$, $u$, $H$, $b$, $Q$, and $R$) and 2) the definition of a mapping from position in the syllable (relative frame time) to the distribution region. Thus, the parameters $F(\alpha_k)$ and $u(\alpha_k, \pi_k)$ of the state equation and $H(\beta_k, \gamma_k)$ of the observation equation are conditioned on the region within the syllable to capture timing effects (e.g., to model peak timing within the syllable) in addition to the previously mentioned dependencies. The number of regions is determined empirically.

For synthesis applications, we need a deterministic mapping from observation times to regions, and this work has explored four alternatives:

1) mapping linearly in time across the syllable;
2) mapping linearly in time outward from the center point of the vowel;
3) mapping linearly in normalized time;[4]
4) using a set of mappings specified with a heuristic function of syllable structure.

All of the mappings produce similar results (with differences under ten percent in RMS error), but slightly better results are obtained with mapping linearly outward from the center of the vowel, which is therefore used in the results reported later. The RMS error for log energy is not sensitive to the number of regions and RMS error for $F_0$ versus the number of regions has a shallow minimum occurring at approximately six regions.

### D. Pitch Range

High level (discourse) range and baseline effects on $F_0$ and the log energy are represented by scaling the overall contour at the intermediate prosodic phrase level:

$$F_0 = A\tilde{F}_0 + F_0^b$$

where $\tilde{F}_0$ is a normalized contour generated by the dynamical system model, $F_0^b$ represents the baseline, and $A$ is the range predicted by a regression tree based on simple "discourse" features, as described below. In fact, we use $A = F_0^p - F_0^b$, where $F_0^p$ is the peak $F_0$ value in the intermediate phrase. The baseline term $F_0^b$ could also be predicted, but we could not find features to predict it reliably and used a constant value of 68 Hz, the mean detected baseline value for this speaker. (Energy can be normalized similarly.)

Range scaling impacts training of the dynamical system model as well as the regression tree. In each intermediate phrase, the peak $F_0$ is found by picking the median filtered $F_0$ value (filter length of five frames for a window length of 50 ms) corresponding to the peak energy during the vowel in each accented syllable in the phrase, and then choosing the maximum of this set. Median filtering is used only in estimating the pitch range, to eliminate false peaks caused by $F_0$ tracking errors; the raw $F_0$ waveform is used in the rest of the model where we explicitly represent segmental effects. The $F_0$ peaks were used for normalizing data

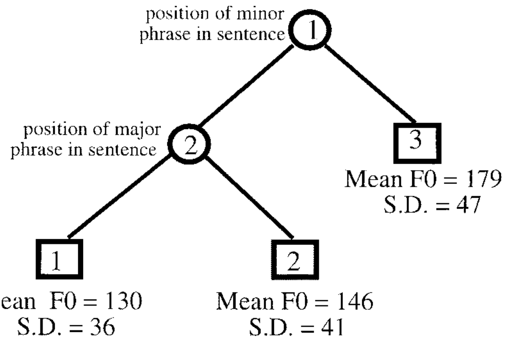

position of minor phrase in sentence
1
position of major phrase in sentence
2
3
Mean F0 = 179
S.D. = 47
1
Mean F0 = 130
S.D. = 36
2
Mean F0 = 146
S.D. = 41

Fig. 3. Regression tree for predicting the peak $F_0$ for an intermediate phrase.

$$\tilde{F}_0 = \frac{F_0 - F_0^b}{F_0^p - F_0^b} \tag{7}$$

for training the dynamical system model. The peak minus baseline value, $A = F_0^p - F_0^b$, was used for training the regression tree, which was designed with the standard algorithm developed by Breiman *et al.* [34].

In the work here, we considered only very simple characterizations of discourse structure as tree questions, including minor and major phrase positions within sentences and paragraphs, phrase lengths, and the number of pitch accents in the phrase. The resultant tree uses only features that indicate where the current phrase falls in its sentence, where a higher peak $F_0$ is predicted for phrases that occur early in the sentence, as shown in Fig. 3. The tree gave RMS error of 42.1 Hz on the test set, compared to a variance of 45.7 Hz on that data for a speaker with a mean peak $F_0$ of 215 Hz. The reduction in variance is probably small because of the simplistic prediction variables used, and could be significantly improved with access to more sophisticated discourse analysis, of the sort that would be available in a message-to-speech application.

## IV. Parameter Estimation

This section goes into more detail the parameter estimation equations needed in the two-step iterative algorithm, first giving a modification of that developed in Digalakis *et al.* [8] that includes nonzero means in both state and observation equations and then expanding on the solution to handle partial observations (needed for unvoiced regions) and nonuniform parameter tying.

### A. Baseline Algorithm

Finding the dynamical system parameters automatically with traditional techniques is difficult, particularly since there are missing observations in the unvoiced regions. Therefore,

[4] Linear mapping in normalized time involves first warping the observations to a new time scale based on normalized duration, where the normalized duration of a phone is the observed duration minus the phone-dependent mean duration divided by the corresponding standard deviation. Observations in this time scale map linearly to regions in the model.

this work uses an iterative method for parameter estimation based upon an algorithm developed by Digalakis *et al.* [8] for speech recognition. This approach to parameter estimation uses the two-step, iterative expectation-maximization (EM) algorithm [31] for maximum likelihood (ML) estimation of processes with unobserved components. In this case, the unobserved components are the state vectors, $x_k$, and the $F_0$ component of $y_k$ for unvoiced data.

The EM algorithm is based on the principle that the likelihood of the observed data $\mathcal{L}(\mathbf{Y}, \theta)$ can be maximized by iteratively maximizing the expected joint likelihood of the observed and missing data $E[\mathcal{L}(\mathbf{X}, \mathbf{Y}, \theta) \mid \mathbf{Y}]$, where for the dynamical system model

$$\begin{aligned}\mathcal{L}(\mathbf{X},\mathbf{Y},\theta) = &-\sum_k \{\log|Q_j| + (x_{k+1} - F_j x_k - u_j)^T \\ &\times Q_j^{-1}(x_{k+1} - F_j x_k - u_j)\} \\ &- \sum_k \{\log|R_j| + (y_k - H_j x_k - b_j)^T \\ &\times R_j^{-1}(y_k - H_j x_k - b_j)\} + \text{constant} \qquad (8)\end{aligned}$$

where $j$ is short for $\rho(k)$. The EM algorithm involves iterating

1) **E-step:** Compute $E_{\theta(p)}[\mathcal{L}(\mathbf{X}, \mathbf{Y}, \theta) \mid \mathbf{Y}]$
2) **M-step:** Reestimate
$\theta(p+1) = \text{argmax}_\theta E_{\theta(p)}[\mathcal{L}(\mathbf{X},\ \mathbf{Y}, \theta) \mid \mathbf{Y}]$

until convergence. For the dynamical system model described here, the M-step involves multivariate regression using first and second-order conditional expectations of $x_k$ and $y_k$, which are thus the only computations needed in the E-step. The equations required for these two steps at each iteration $p$ are summarized below for the case where the parameters in region $j$, $\theta_j = \{F_j, u_j, Q_j, H_j, b_j, R_j\}$ are constant for all $k \in \kappa_j = \{k : \rho(k) = j\}$ and parameters are not tied across regions. (We use the term “region” since a single model may have multiple regions to capture variation in time.) Training with the EM algorithm typically converges quickly, taking four to ten iterations in the experiments reported here.

*1) E-Step:* The first- and second-order moments of $y_k$ and $x_k$ for all $k \in \kappa_j$ are needed for parameter estimation, as summarized below. Assuming, for the moment, that $y_k$ is completely observed, these moments simplify to

$$E_{\theta(i)}[x_k \mid \mathbf{Y}] = \hat{x}_{k|N} \qquad (9)$$

$$E_{\theta(i)}\left[y_k x_k^T \mid \mathbf{Y}\right] = y_k \hat{x}_{k|N} \qquad (10)$$

$$E_{\theta(i)}\left[y_k y_k^T \mid \mathbf{Y}\right] = y_k y_k^T \qquad (11)$$

$$E_{\theta(i)}\left[x_k x_k^T \mid \mathbf{Y}\right] = \hat{x}_{k|N} \hat{x}_{k|N}^T + \Sigma_{k|N} \qquad (12)$$

$$E_{\theta(i)}\left[x_k x_{k-1}^T \mid \mathbf{Y}\right] = \hat{x}_{k|N} \hat{x}_{k-1|N}^T + \Sigma_{k,k-1|N} \qquad (13)$$

where $\Sigma_{k,k-1|N} = E_{\theta(i)}[(x_k - \hat{x}_{k|N})(x_{k-1} - \hat{x}_{k-1|N})^T \mid \mathbf{Y}]$ and $\Sigma_{k|N} \equiv \Sigma_{k,k|N}$ is defined similarly. (Without loss of generality, we use $k = 0, \cdots, N$ to represent the times in a contiguous subsequence of $k$ that correspond to region $j$.) Thus, the E-step simply involves computing the quantities $\hat{x}_{k|N}$, $\Sigma_{k|N}$, and $\Sigma_{k,k-1|N}$, which are the expected moments of $x_k$ given the observed data $\mathbf{Y}$. These quantities are computed recursively, as summarized below. All of the recursion equations are standard equations for the Rauch–Tung–Striebel algorithm [32], using the standard Kalman filtering forward recursions followed by a backward pass, except for the cross covariance equations (19) and (20), which were derived by Digalakis *et al.* [8].

*E-Step—Forward Recursions:*

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k e_k \qquad (14)$$

$$\hat{x}_{k+1|k} = F\hat{x}_{k|k} + u \qquad (15)$$

$$e_k = y_k - b - H\hat{x}_{k|k-1} \qquad (16)$$

$$K_k = \Sigma_{k|k-1} H^T (H \Sigma_{k|k-1} H^T + R)^{-1} \qquad (17)$$

$$\Sigma_{k|k} = (I - K_k H)\Sigma_{k|k-1}(I - K_k H)^T + K_k R K_k^T \qquad (18)$$

$$\Sigma_{k+1,k|k+1} = (I - K_{k+1} H) F \Sigma_{k|k} \qquad (19)$$

$$\Sigma_{k+1|k} = F \Sigma_{k|k} F^T + Q \qquad (20)$$

*E-Step—Backward Recursions:*

$$\hat{x}_{k-1|N} = \hat{x}_{k-1|k-1} + A_{k-1}(\hat{x}_{k|N} - \hat{x}_{k|k-1}) \qquad (21)$$

$$\Sigma_{k-1|N} = \Sigma_{k-1|k-1} + A_{k-1}(\Sigma_{k|N} - \Sigma_{k|k-1}) A_{k-1}^T \qquad (22)$$

$$A_{k-1} = \Sigma_{k-1|k-1} F^T \Sigma_{k|k-1}^{-1} \qquad (23)$$

$$\Sigma_{k,k-1|N} = \Sigma_{k,k-1|k} + (\Sigma_{k|N} - \Sigma_{k|k})\Sigma_{k|k}^{-1} \Sigma_{k,k-1|k} \qquad (24)$$

*2) M-Step:* In normal ML estimation, $\hat{\theta}$ would be chosen to maximize the log-likelihood of the data, but because $\mathbf{X}$ is unobserved and $\mathbf{Y}$ is only partially observed, $\hat{\theta}$ is chosen to maximize the conditional expected value of the log-likelihood given the old value for $\theta$. The solutions for the dynamical system model parameters, (25)–(28), shown at the bottom of the next page, are therefore specified in terms of conditional expected values calculated during the E-step. where $\kappa_j' \subset \kappa_j$ including all “$k < N$” in a run and $|A|$ represents the number of elements in $A$. The initial state mean $\hat{\mu}_0$ and covariance $\hat{\Sigma}_0$ are given by the standard ML estimates taking the average over $\{\hat{x}_{0|N}\}$ and $\{\Sigma_{0|N}\}$, sets that include moments for the initial state in each contiguous subsequence that maps to model $j$.

## *B. Handling Unobserved Regions*

Part of the advantage of using the EM algorithm for parameter estimation is its ability to handle missing observations. Unlike the speech recognition application considered in [8], where whole vectors $y_k$ might be missing, the joint $(F_0, E)$ modeling application has the problem that only part of $y_k$ may be missing. Specifically, $F_0$ observations are missing during unvoiced regions of the speech and during periods where $F_0$ data is ignored due to $F_0$ tracking errors. For such times $k$, the statistics $E_{\theta(i)}[y_k \mid \mathbf{y}]$, $E_{\theta(i)}[y_k x_k^T \mid \mathbf{Y}]$ and $E_{\theta(i)}[y_k y_k^T \mid \mathbf{Y}]$ must also be computed in the E-step.

Recall that $y_k = H x_k + b + v_k$, where $v_k$ is a Gaussian process with zero mean and covariance $R$ that is independent of $x_k$. The distribution of $x_k$ is normal with mean $\hat{x}_{k|N}$ and covariance $\Sigma_{k|N}$. Then the distribution of $y_k$ is also normal, with mean and covariance

$$\mu_k = H\hat{x}_{k|N} + b \tag{29}$$

$$\xi_k = H\Sigma_{k|N}H^T + R. \tag{30}$$

Let the observation vector be represented by $y_k = (y_k^1 \;\; y_k^2)^T$ where $y_k^1$ is the $F_0$ component and $y_k^2$ is the energy component. Similarly, each of the other model parameters are represented by their matrix coordinates in the following way:

$$x_k = \begin{pmatrix} x_k^1 \\ x_k^2 \end{pmatrix} \quad \xi_k = \begin{pmatrix} \xi_k^{11} & \xi_k^{12} \\ \xi_k^{21} & \xi_k^{22} \end{pmatrix}.$$

To compute the additional expectations, we use the fact that $p(y_k^1 \mid y_k^2)$ and $p(y_k, x_k)$ are Gaussian with

$$E_{\theta(i)}\left[y_k^1 \,\middle|\, \mathbf{Y}\right] = \mu_k^1 + \left(y_k^2 - \mu_k^2\right)\xi_k^{12}/\xi_k^{22} \tag{31}$$

$$E_{\theta(i)}\left[y_k^1 y_k^1 \,\middle|\, \mathbf{Y}\right] = \xi_k^{11} - \left(\xi_k^{12}\right)^2 \xi_k^{22} + \left(E_{\theta(i)}\left[y_k^1 \,\middle|\, \mathbf{Y}\right]\right)^2 \tag{32}$$

$$\begin{aligned} E_{\theta(i)}\left[y_k^1 x_k^1 \,\middle|\, \mathbf{Y}\right] = {} & H^{11}E_{\theta(i)}\left[x_k^1 x_k^1 \,\middle|\, \mathbf{Y}\right] \\ & + H^{12}E_{\theta(i)}\left[x_k^1 x_k^2 \,\middle|\, \mathbf{Y}\right] + b^1 E_{\theta(i)}\left[x_k^1 \,\middle|\, \mathbf{Y}\right] \end{aligned} \tag{33}$$

$$\begin{aligned} E_{\theta(i)}\left[y_k^1 x_k^2 \,\middle|\, \mathbf{Y}\right] = {} & H^{11}E_{\theta(i)}\left[x_k^1 x_k^2 \,\middle|\, \mathbf{Y}\right] \\ & + H^{12}E_{\theta(i)}\left[x_k^2 x_k^2 \,\middle|\, \mathbf{Y}\right] + b^1 E_{\theta(i)}\left[x_k^2 \,\middle|\, \mathbf{Y}\right] \end{aligned} \tag{34}$$

where $E_{\theta(i)}[x_k^j \mid \mathbf{Y}]$ is the $j$th element of $\hat{x}_{k|N}$.[5] These terms are then used in the modified conditional expectations, which replace (10) and (11), as follows:

$$E_{\theta(i)}[y_k \mid \mathbf{Y}] = \begin{cases} y_k, & \text{if observed} \\ \begin{pmatrix} E_{\theta(i)}\left[y_k^1 \,\middle|\, \mathbf{Y}\right] \\ y_k^2 \end{pmatrix}, & \text{if missing} \end{cases} \tag{35}$$

$$\begin{aligned} & E_{\theta(i)}\left[y_k y_k^T \,\middle|\, \mathbf{Y}\right] \\ & = \begin{cases} y_k y_k^T, & \text{if observed} \\ \begin{pmatrix} E_{\theta(i)}\left[y_k^1 y_k^1 \,\middle|\, \mathbf{Y}\right] & E_{\theta(i)}\left[y_k^1 \,\middle|\, \mathbf{Y}\right] y_k^2 \\ y_k^2 E_{\theta(i)}\left[y_k^1 \,\middle|\, \mathbf{Y}\right] & y_k^2 y_k^2 \end{pmatrix}, & \text{if missing} \end{cases} \end{aligned} \tag{36}$$

$$\begin{aligned} & E_{\theta(i)}\left[y_k x_k^T \,\middle|\, \mathbf{Y}\right] \\ & = \begin{cases} y_k E_{\theta(i)}\left[x_k^T \,\middle|\, \mathbf{Y}\right], & \text{if observed} \\ \begin{pmatrix} E_{\theta(i)}\left[y_k^1 x_k^1 \,\middle|\, \mathbf{Y}\right] & E_{\theta(i)}\left[y_k^1 x_k^2 \,\middle|\, \mathbf{Y}\right] \\ y_k^2 E_{\theta(i)}\left[x_k^1 \,\middle|\, \mathbf{Y}\right] & y_k^2 E_{\theta(i)}\left[x_k^2 \,\middle|\, \mathbf{Y}\right] \end{pmatrix}, & \text{if missing.} \end{cases} \end{aligned} \tag{37}$$

When the component $y_k^1$ of the observation is missing there are also some minor changes to the other calculations for the E-step: the filtered estimate $\hat{x}_{k|k}$ and its variance $\Sigma_{k|k}$ are calculated using the value for $y_k^1$ resulting from the predicted estimate $\hat{x}_{k|k-1}$. This produces an innovations term [see (16)] such that $e_k^1 = 0$ and the update is based entirely upon the observed component of $y_k$. The equation for calculating the Kalman gain of (17) does not change, nor are the backward recursions affected by missing observations since they do not depend on the observed process $\mathbf{Y}$.

### *C. Parameter Tying*

In the general form for the dynamical system model specified in (1) and (2), the subscript $j$ identifies the model region and thus the set of parameters appropriate for $y_k$ and $x_k$, where $j = \rho(k)$. If two regions, say $j_1$ and $j_2$, share some but not all parameters, then the optimal reestimation equations are more complex than those given in the M-step in Section IV-A. In particular, the solution becomes more difficult when $H$ and $b$ (or $F$ and $u$) have different parameter dependencies.

For a true *ML solution*, the M-step equations must be modified to simultaneously solve for all interdependent parameters. For example, if the parameters $(H_j, b_j) \in \theta_j$ in the state equation [see (1)] are tied differently across regions $j$, then these parameters must be solved for jointly with those

$$[\hat{H}_j \quad \hat{b}_j] = \left[\sum_{k\in\kappa_j} E\left[y_k x_k^T \,\middle|\, \mathbf{Y}\right] \quad \sum_{k\in\kappa_j} E[y_k \mid \mathbf{Y}]\right] \times \begin{bmatrix} \sum\limits_{k\in\kappa_j} E\left[x_k x_k^T \,\middle|\, \mathbf{Y}\right] & \sum\limits_{k\in\kappa_j} E[x_k \mid \mathbf{Y}] \\ \sum\limits_{k\in\kappa_j} E\left[x_k^T \,\middle|\, \mathbf{Y}\right] & 1 \end{bmatrix}^{-1} \tag{25}$$

$$[\hat{F}_j \quad \hat{u}_j] = \left[\sum_{k\in\kappa_j{}'} E\left[x_{k+1} x_k^T \,\middle|\, \mathbf{Y}\right] \quad \sum_{k\in\kappa_j{}'} E[x_{k+1} \mid \mathbf{Y}]\right] \times \begin{bmatrix} \sum\limits_{k\in\kappa_j{}'} E\left[x_k x_k^T \,\middle|\, \mathbf{Y}\right] & \sum\limits_{k\in\kappa_j{}'} E[x_k \mid \mathbf{Y}] \\ \sum\limits_{k\in\kappa_j{}'} E\left[x_k^T \,\middle|\, \mathbf{Y}\right] & 1 \end{bmatrix}^{-1} \tag{26}$$

$$\hat{R}_j = \frac{1}{|\kappa_j|}\sum_{k\in\kappa_j} E\left[y_k y_k^T \,\middle|\, \mathbf{Y}\right] - \left[\frac{1}{|\kappa_j|}\sum_{k\in\kappa_j} E\left[y_k x_k^T \,\middle|\, \mathbf{Y}\right] \quad \frac{1}{|\kappa_j|}\sum_{k\in\kappa_j} E[y_k \mid \mathbf{Y}]\right] \times [\hat{H} \quad \hat{b}]^T \tag{27}$$

$$\hat{Q} = \frac{1}{|\kappa_j'|}\sum_{k\in\kappa_j{}'} E\left[x_{k+1} x_{k+1}^T \,\middle|\, \mathbf{Y}\right] - \left[\frac{1}{|\kappa_j'|}\sum_{k\in\kappa_j{}'} E\left[x_{k+1} x_k^T \,\middle|\, \mathbf{Y}\right] \quad \frac{1}{|\kappa_j'|}\sum_{k\in\kappa_j'} E[x_{k+1} \mid \mathbf{Y}]\right] \times [\hat{F} \quad \hat{u}]^T \tag{28}$$

[5] These equations simplify if $y_k^1$ and $y_k^2$ are assumed independent, in which case $H$, $R$ and $\Sigma_{k|N}$ are diagonal.

from all regions that either one is tied to. The same is true for $(F_j, u_j)$ from the observation equation [see (2)]. The parameter tying problem is handled by thinking of the matrix derivatives in terms of vector derivatives with respect to one row at a time. Then these vectors are stacked up, combining the appropriate conditional expectations from all the $H_j$'s and $b_j$'s (for example) that share the same subset of the data. The solution is then an extended version of (25) and (26), requiring a potentially large matrix inversion.

For example, consider the following relatively simple case with parameter tying only in the observation equation:

$$y_k = H(\beta_k)x_k + b(\beta_k, \gamma_k) + v_k. \tag{38}$$

In this case, both $H$ and $b$ have the data broken up into regions according to the factor $\beta_k$, but for $b$ the data must be further partitioned according to the factor $\gamma_k$. If $\gamma_k$ can take on two values ($\gamma^1$ and $\gamma^2$) then the data is broken up to define two parameters $b(\beta, \gamma^1)$ and $b(\beta, \gamma^2)$ for every $H(\beta)$. Let $\kappa_{\beta,i} = \{k : \beta_k = \beta, \gamma_k = \gamma^i\}$ and $\kappa_\beta = \kappa_{\beta,1} \cup \kappa_{\beta,2}$. The new equations for the parameters $H$ and $b$ in terms of the conditional expected values are in (39), shown at the bottom of the page.

If there are only a small number of cross dependencies, then this optimal solution is appropriate. However, extensive cross dependencies will lead to a large matrix inversion problem, which requires substantial training data for robust estimation and computational power for inversion.

When the optimal solution described above is impractical, an *approximate solution* for the parameters can be used. Using the observation equation as an example, our approach is to estimate $H_j$ assuming that $b_j$ is known, taking it from the previous EM iteration, and then reestimate $b_j$ and $R_j$ using this estimate of $H_j$. The solution for the above example is

$$\hat{H}(\beta)^{(p+1)} = \left(\sum_i \sum_{k\in\kappa_{\beta,i}} E\big[(y_k - b^{(p)}_{\beta,i})x_k^T \,\big|\, \mathbf{Y}\big]\right) \times \left(\sum_{k\in\kappa_\beta} E\big[x_k x_k^T \,\big|\, \mathbf{Y}\big]\right)^{-1} \tag{40}$$

$$\hat{b}^{(p+1)}_{\beta,i} = \frac{1}{|\kappa_{\beta,i}|} \sum_{k\in\kappa_{\beta,i}} \left(E[y_k \mid \mathbf{Y}] - \hat{H}(\beta)^{(p+1)}\hat{x}_{k|N}\right) \tag{41}$$

where $p$ indicates the EM iteration number. The solution for the state equation is analogous. The derivation is a variation of that in [8], details of which can be found in [33]. The expected moments remain the same as given in previous sections.

## V. Experimental Results

This section presents results for $F_0$ and log energy contour generation using the dynamical system model with the approximate ML estimation technique. The results given here are in the form of quantitative measures as well as perceptual evaluations of the $F_0$ generation algorithms.

### A. Radio News Corpus

The experiments are based on utterances that were collected from a single female radio announcer (F2B) from the Boston University radio news corpus [35]. We used 34 stories (approximately 48 min, 8841 words) recorded from actual radio broadcasts for model training, and four stories recorded in the laboratory (approximately 11 min) for testing. Lexical stress assignments and syllable boundaries were taken from a large dictionary, and phonetic labels and segment boundaries were obtained automatically using the Boston University speech recognition system. Energy and $F_0$ contours were computed with the Entropic Waves version 5.0 pitch tracker [36] every 10 ms. About 250 longer duration pitch tracking errors (more than 50 ms) were hand marked to allow the programs to treat these values as missing during training and testing.

Prosodic labels for this work are based upon the ToBI [20]–[22] multitiered labeling system. The break index tier specifies prosodic constituent structure using a subset of the markers in [17], and the tone tier contains a subset of Pierrehumbert's tone labels [30], [13]. Every intermediate phrase (break index three or four) is assigned either an "H-," "!H-" or "L-" marker corresponding to a final high, downstepped or low tone. An intonational phrase (break index 4) has a final boundary tone marked by either a "L%" or an "H%." Since intonational phrases are composed of one or more intermediate phrases plus a boundary tone, full intonation phrase boundaries will have two final tones, e.g., "L-L%." Pitch accent tones are marked at every accented syllable, or about one third of the syllables in this corpus. Accent types include: a peak accent "H*," a low accent "L*," a scooped accent "L*+H," a rising peak accent "L+H*," and a down-stepped peak "H+!H*." Down-stepped high tones (a high tone after a higher tone in

$$[\hat{H}(\beta) \quad \hat{b}(\beta,\gamma^1) \quad \hat{b}(\beta,\gamma^2)] = \left[\sum_{k\in\kappa_\beta} E\big[y_k x_k^T \,\big|\, \mathbf{Y}\big] \quad \sum_{k\in\kappa_{\beta,1}} E[y_k \mid \mathbf{Y}] \quad \sum_{k\in\kappa_{\beta,2}} E[y_k \mid \mathbf{Y}]\right] \times \begin{bmatrix} \sum_{k\in\kappa_\beta} E\big[x_k x_k^T \,\big|\, \mathbf{Y}\big] & \sum_{k\in\kappa_{\beta,1}} E[x_k \mid \mathbf{Y}] & \sum_{k\in\kappa_{\beta,2}} E[x_k \mid \mathbf{Y}] \\ \sum_{k\in\kappa_{\beta,1}} E\big[x_k^T \,\big|\, \mathbf{Y}\big] & 1 & 1 \\ \sum_{k\in\kappa_{\beta,2}} E\big[x_k^T \,\big|\, \mathbf{Y}\big] & 1 & 1 \end{bmatrix}^{-1} \tag{39}$$

TABLE I
NUMBER OF TRAINING TOKENS FOR THE DIFFERENT PROSODIC LABEL CLASSES

| Accent | Phrase Boundary Markers | | | | | |
|---|---|---|---|---|---|---|
| Labels | none | L-L% | L-H% | L- | H- | !H- |
| unmarked | 5566 | 670 | 385 | 182 | 120 | 62 |
| stressed | 2254 | 215 | 88 | 56 | 21 | 31 |
| H* | 2886 | 186 | 129 | 33 | 115 | 46 |
| ! H* | 785 | 193 | 71 | 21 | 42 | 20 |
| L* | 150 | 7 | 38 | 3 | 2 | 0 |

the same intermediate phrase) are marked by a "!." Uncertainty about a phrase tone or pitch accent can be marked with a "?," but was used by the transcribers on less than 2% of the accents marked.

Prosodic labeling for these stories was done at two sites, and three of the stories containing 1002 words were labeled by both sites so that consistency of the labels could be assessed [33].[6] For pitch accent types, there was agreement on presence versus absence for 91% of the words. On those 487 words that were marked by both labeling groups with an accent, there was 60% agreement on accent type with most of the disagreements occurring for the difficult "L+H*" versus "H*" distinction. Grouping "H*" tones with "L+H*" tones as in Pitrelli *et al.* [21], there was 81% agreement on pitch accent type. Boundary tone agreement was 93% for the 207 words marked by both labelers with an intonational phrase boundary, and similarly there was 91% agreement for 280 phrase accents. Similar levels of agreement were found for labeling the five ToBI break index levels.

The model developed in this paper uses five categories of prominence labels for syllables ("unmarked" and "stressed" for unaccented syllables, and "high," "downstepped high" and "low" for accents) and six categories of phrase tones ("none," "L–L%," "L–H%," "L–," "H–," and "!H–"). Different accent types were grouped as needed to get high labeling accuracy (e.g., "H*" grouped with "L+H*") and/or sufficient training data (e.g., "X*"? grouped with downstepped high, and "*?" with stressed). The division of the unaccented category into unmarked and stressed syllables was based on dictionary lexical stress and motivated by energy modeling. The number of tokens in the training set for the resulting 30 possible combinations are summarized in Table I.

### B. Numerical Results

Experiments with the $F_0$ model were run using several different configurations, in an attempt to separately evaluate the different components of the model (contour modeling versus range prediction) and to assess various modeling assumptions. Performance was measured in terms of RMS error between predicted $F_0$ and energy contours and observed contours in the test data, using the ToBI prosodic labels associated with the test data as input for controlling parameter selection.

TABLE II
RMS ERRORS FOR *NORMALIZED* $F_0$ AND LOG ENERGY FOR CONTOURS REPRESENTED BY THE DYNAMICAL SYSTEM MODEL ON TRAINING AND TEST DATA. NOTE THAT THE RANGE OF THE NORMALIZED FEATURES IS $[0, 1]$

| | $F_0$ | | Energy | |
|---|---|---|---|---|
| | Training | Test | Training | Test |
| Detected baseline | 0.198 | 0.222 | 0.171 | 0.217 |
| Constant baseline | 0.199 | 0.199 | 0.194 | 0.210 |

TABLE III
RMS ERRORS FOR $F_0$ (Hz) AND LOG RMS ENERGY (dB) FOR DYNAMICAL SYSTEM MODEL WITH DIFFERENT LEVELS OF THE MODEL AUTOMATED COMPARED TO VARIANCE OF THE TEST AND TRAINING DATA

| | $F_0$ (Hz) | | Energy (dB) | |
|---|---|---|---|---|
| | Training | Test | Training | Test |
| Joint predicted contour, true range | 29.2 | 28.5 | 2.68 | 3.06 |
| Indep predicted contours, true range | 29.4 | 28.7 | 2.73 | 3.06 |
| Joint predicted contours, predicted range | 34.2 | 34.7 | 2.94 | 3.48 |
| Variance | 44.2 | 42.8 | 5.55 | 4.88 |

The first experiments look at performance of the dynamical system model separately from the range prediction algorithm. The results for the RMS errors for the dynamical system model for range-normalized $F_0$ and energy contours are shown in Table II. Normalization is based on the detected peak for the phrase, as described in Section III-D. Results are given for the case where the baseline is detected separately for each phrase versus a constant baseline for all phrases in all utterances. The detected baseline is the observed minimum for the phrase, with some corrections for $F_0$ tracking errors. The use of a constant baseline is motivated by the difficulty in automatically predicting the baseline from paragraph structure, but the results here suggest that baseline prediction (even if perfect) would not improve results on the test data given a definition of baseline as the minimum value in the phrase. The corresponding RMS errors of the constant-baseline model output, scaled by the detected peaks, are given in the first line of Table III.

The next experiment compares the joint vector model of $F_0$ and energy to independent scalar models. Comparing the first and second rows of Table III shows that there is a small reduction in $F_0$ error associated with the joint model (not significant), and none for energy. Since the independent models are simpler to implement, the joint model is probably unnecessary for synthesis of this particular speaking style.

Finally, comparing rows one and three of Table III shows some degradation in performance associated with using the automatic pitch range prediction algorithm versus the true peak for the phrase. The variance for the $F_0$ (Hz) and log energy (dB) of the training and test data are shown in the last row of Table III, for reference. The results show that, with an observation-based $F_0$ range predictor, the dynamical system model can reduce variance by 33% for $F_0$ and 37% for energy. With automatic range prediction, the gains are smaller: 19% for $F_0$ and 29% for energy. Note that range is a much more important factor for $F_0$ than for energy, with a larger loss in

[6] For these three stories, we arbitrarily chose one version for use in model training.

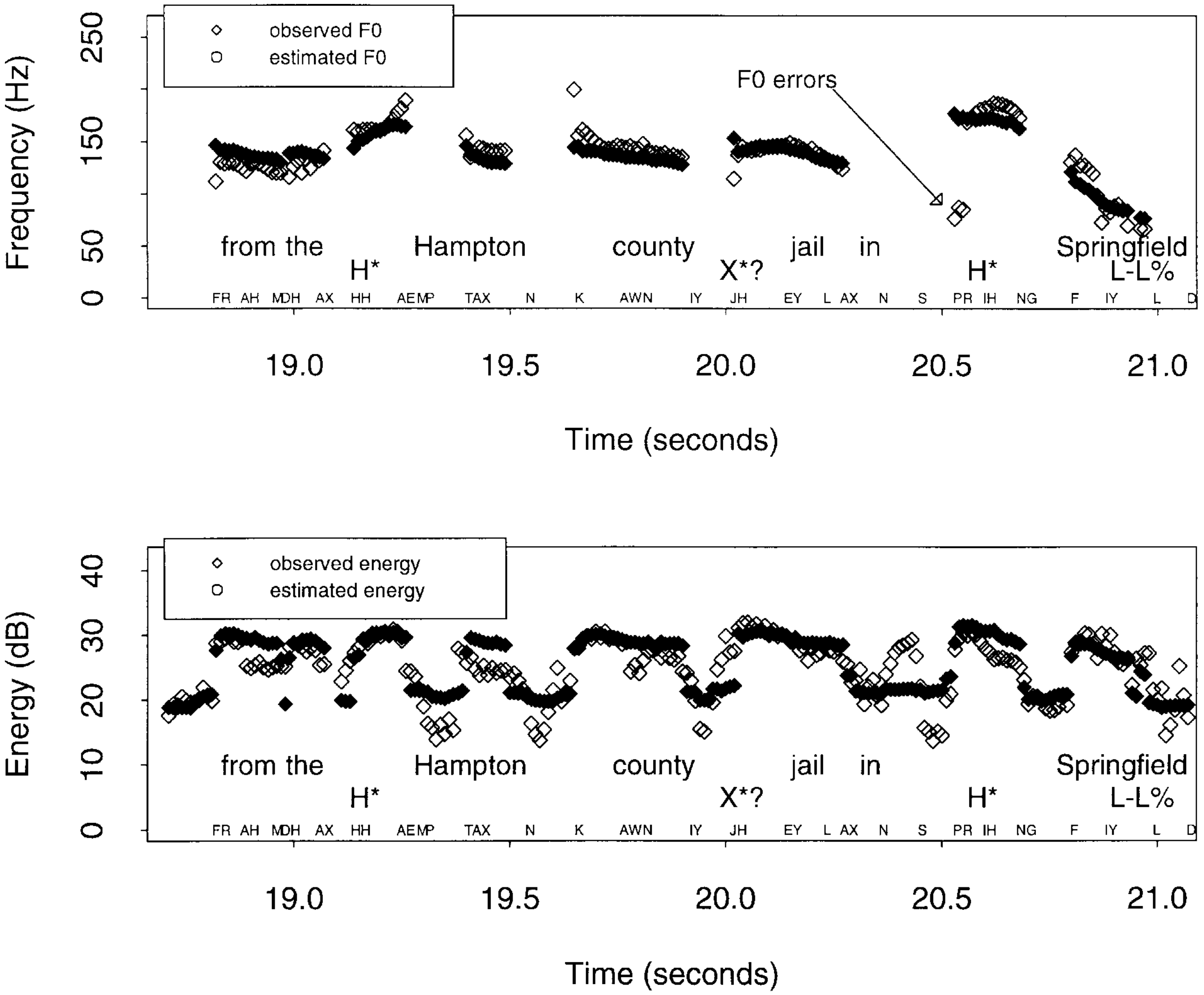


Fig. 4. Plot showing a sample of observed and estimated $F_0$ contours for a phrase from the test data set. Phone labels are placed at the end of the respective segment.

performance for predicted $F_0$ range than for predicted energy, despite the fact that predicted energy range is constant for all phrases.

Several additional experiments were conducted to assess some of the parameter dependence choices. The changes in RMS error were relatively small (less than 10%), but it is interesting to note which changes did not hurt performance. Changing the accent categories from high, downstepped high, and low, to just the presence of an accent increased the RMS errors for $F_0$ but did not affect the RMS errors for energy. Similar results were observed when the number of phrase tone categories was reduced from five to one. Removing all segmental effects (dependence on $\gamma_k$) increased the RMS errors for both energy and $F_0$. The additive term $b^1(\beta_k)$ (that would capture phrase-final $F_0$ lowering) did not benefit either $F_0$ or energy. (This result does not completely rule out the possibility of phrase-final lowering, since it may be that the three category distinction for $\beta_k$ was too simple.) Completely removing the dependency on phrase position from $H(\beta_k, \gamma_k)$ did not affect the RMS errors for energy and increased the RMS errors for $F_0$ only slightly.

Although the reduction in variance is not as large as we had expected, the model appears to be working well from perceptual experiments, as described in the next section. Looking at sample $F_0$ contours, such as those shown in Fig. 4, we see that some of the large errors are due to pitch tracking errors in the original contour that we do not want the model to predict. Similarly, phrase-final creak also leads to large variations in the energy contour, that should be predictable only in association with predicting creak. However, these plots also illustrate points where the model can be improved.

In $F_0$ prediction, some of the biggest errors appear to be coming from segmental effects, which might be addressed by adding region dependence (i.e., position in the syllable) and an indicator of whether the syllable or phoneme is foot- or phrase-initial. For example, in Fig. 5, the /b/ in "wristband" is the onset consonant of a stressed, foot-initial syllable, and therefore might be expected to be produced with more articulatory precision and have a larger effect on $F_0$ than the /b/ in other environments (e.g., syllable-final or prevocalic in a reduced syllable). Of course, it may also be useful to increase the number of phoneme classes. In addition, it would be useful to have better segmental alignments in the corpus, particularly at vowel-nasal boundaries, because segmental effects are greatest near phoneme boundaries. There is also a large region of error in the $F_0$ contour during the H* accent

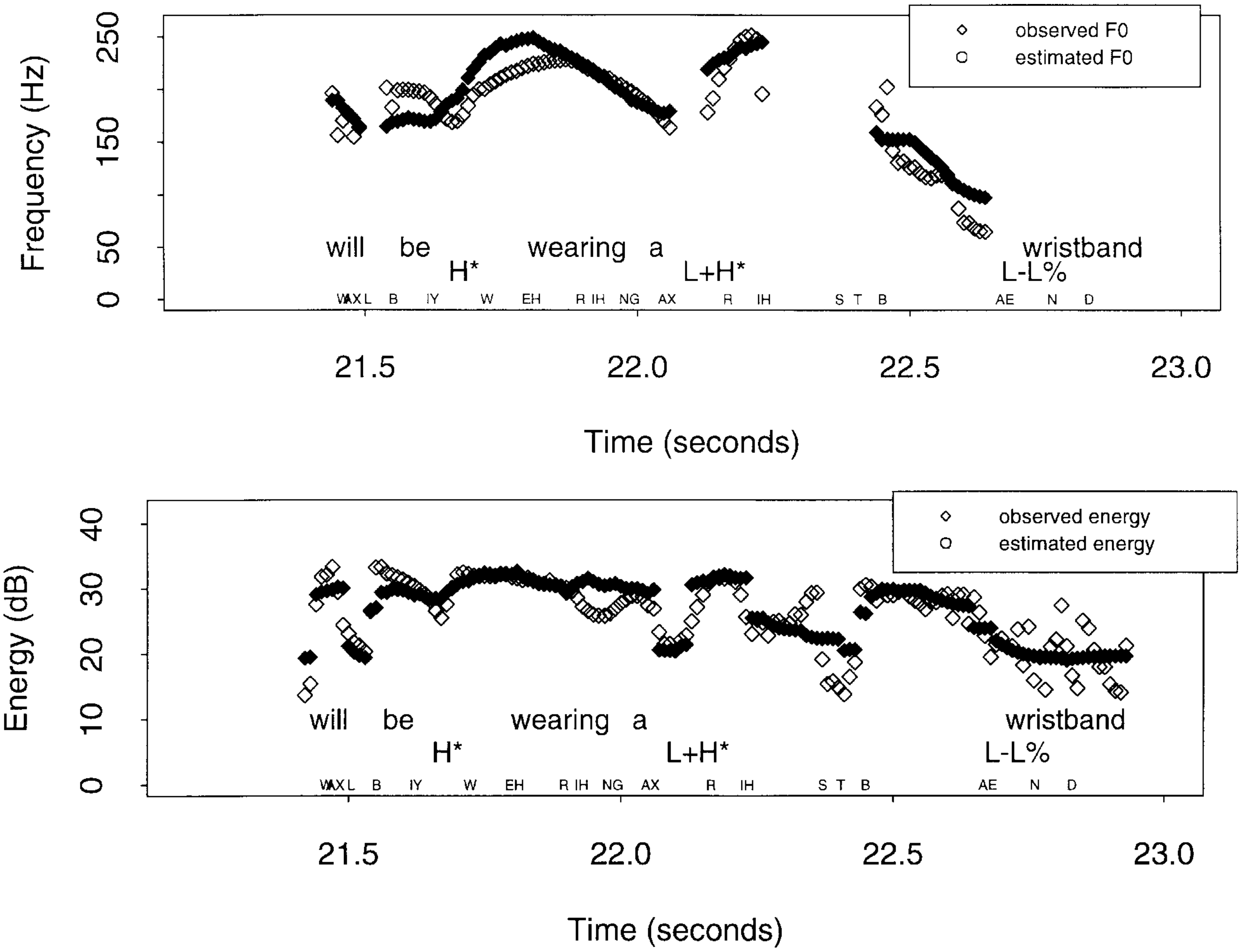


Fig. 5. Plot showing a sample of observed and estimated $F_0$ and log energy contours for a phrase from the test data set. Phone labels are placed at the end of the respective segment.

on "wearing." This might be addressed by improved models of timing within the syllable, or it may be a problem of poor modeling of prominence relationships among accents within the phrase. In this example, the second accent is higher than the first in the original contour, and this sort of prominence boosting is relatively rare in this corpus.

Both Figs. 4 and 5 show some larger errors in the predicted log energy contours in the unvoiced regions of the phrase. The majority of the problems with energy modeling occur during periods of strong frication, during both strong fricatives (e.g., /$s$/) and during the frication phase of the aspiration interval of strong stops. This problem is reduced somewhat relative to experiments that did not have a separate phonetic class for strong fricatives, but the error is still high. Since many of the parameter dependencies were chosen with the $F_0$ contour in mind, and since joint $F_0$ and energy modeling does not seem to be advantageous, it would be useful to explore a different set of parameter dependencies in an independent version of the energy model. In particular, results in [37] suggest that more segmental categories are needed.

Despite remaining problems, the model does seem to be quite good at capturing several characteristics of $F_0$ and energy contours, and the perceptual tests described next show that the $F_0$ model compares favorably to another existing model.

### C. Perceptual Test Results

The $F_0$ model was also tested in two perceptual tests using the 1994 AT&T TTS synthesizer with the female voice. The $F_0$ generation component in the TTS synthesizer is based on the hybrid target/filter model [29], and it provides a high-quality baseline for comparison. The first test examines the dynamical system model's performance when generating only $F_0$ contours, while the second examined the model's performance when generating both $F_0$ and energy contours. For these experiments, the contours generated by the dynamical system model are modified so that the $F_0$ range is on a better scale for the synthesizer, to reduce distortion associated with large $F_0$ changes due to signal processing limitations of concatenative synthesizers. Specifically, the range was linearly reduced to 70% of the original and the baseline was offset by +30 Hz.

Both tests used a similar format, relying on listeners to comparatively rate the naturalness of synthesized speech, comparing three versions of each sentence that varied in the intonation and/or energy contour. (Two tests were run to reduce the amount of listening time asked of a subject in a single experiment.) The test material consisted of 15 sentences from the radio news story test set. All versions were produced using the same word pronunciations and durations from the original spoken versions. The participants listened to the three

TABLE IV
MEAN NATURALNESS RATINGS FOR SPEECH SYNTHESIZED USING DIFFERENT SOURCES FOR THE PROSODIC MARKERS (HAND-LABELED FROM THE ORIGINAL SPEECH VERSUS PREDICTED BY THE SYNTHESIZER) AND THE $F_0$ MODEL (DYNAMICAL SYSTEM VERSUS HYBRID TARGET/FILTER MODEL)

| Prosody Labels | $F_0$ Model | Rating |
|---|---|---|
| Hand-labeled | Dynamical system | 2.5 |
| Hand-labeled | Target/filter | 3.3 |
| Predicted | Target/filter | 3.7 |

versions of each sentence as many times as they wanted and were asked to rate the naturalness on a scale of 1 to 5, with 1 representing the most natural. They were instructed that this test was evaluating several methods of producing intonational contours and that they should pay close attention to the naturalness of the intonation. The order was varied so the participants did not know which version of the synthesized speech they were listening to.

For the first test, the three different versions of each sentence included: 1) our $F_0$ model with hand-labeled prosody, 2) the hybrid target/filter $F_0$ model (TTS default) with prosody labels as marked in the spoken version, and 3) the hybrid target/filter $F_0$ model and predicted prosody labels. The results for this test, based on 12 subjects, are shown in Table IV. Each subject's scores were averaged across the 15 sentences, and the averaged scores were analyzed by means of analysis of variance (ANOVA). Using a one-way analysis of variance, a test to determine if the scores are from the same distribution fails at the .01 level: $F_c = 105.9 > 4.8 > F_{.99}(2, 537)$. A multiway analysis shows significant differences between each pair. Comparing the dynamical system model (mean 2.5) to the target/filter model (mean 3.3), we obtain $p < 10^{-4}$, $t_{11} = 10.6$. A Duncan multiple range test also shows that the differences across the three cases are significant at the .01 level. From a table of studentized ranges, we calculate the least significant range $R_2 = 0.18 < 0.40$ for the ordered mean difference between case one and two, and $R_2 = 0.18 < 0.81$ for the ordered mean difference between case two and three. Interestingly, the difference in performance was greater for the different $F_0$ models than for the different sources of prosody markers, but results show that synthetic speech can be improved both by better $F_0$ modeling and by better prosodic marker prediction.

In the second test, we compared three versions of the same 15 sentences using: 1) the dynamical system $F_0$ and energy model, 2) the dynamical system $F_0$ model and the synthesizer's default energy model, and 3) the target/filter $F_0$ and default energy model. All three versions are synthesized using the hand-labeled prosodic markers. The results for this test, based on eight listeners, confirmed the finding of a significant improvement in the intonation model but did not show a difference between the energy models. The averaged scores were again analyzed using anova and Duncan tests. Using a one-way analysis of variance, a test to determine if the scores are from the same distribution fails at the .01 level: $F_c = 19.88 > 4.8 > F_{.99}(2, 357)$. As shown in Table V, the dynamical system model for $F_0$ scored 2.7 compared to 3.4 for the hybrid target/filter system (0.64 average difference, $p < 0.0025$, $t_7 = 4.5$). Using the dynamical system for the energy model resulted in a slightly worse rating (2.8 versus 2.7) (average difference $= -0.11$, n.s.). A Duncan multiple range test again confirms these results. The least significant range results are $R_2 = 0.28 > 0.11$ for the ordered mean difference between case one and two (not significant) and $R_2 = 0.28 < 0.64$ for the ordered mean difference between case two and three (significant at .01 level). Since the energy model is clearly off in low energy regions, we are optimistic that addressing this issue will lead to improved perceived quality.

TABLE V
MEAN NATURALNESS RATINGS FOR SPEECH SYNTHESIZED USING DIFFERENT SOURCES FOR THE $F_0$ MODEL (DYNAMICAL SYSTEM VERSUS HYBRID TARGET/FILTER MODEL) AND ENERGY MODEL (DYNAMICAL SYSTEM VERSUS TTS DEFAULT). IN ALL CASES, THE PROSODIC MARKERS ARE HAND-LABELED FROM THE ORIGINAL SPEECH

| $F_0$ Model | Energy Model | Rating |
|---|---|---|
| Dynamical system | TTS Default | 2.7 |
| Dynamical system | Target/filter | 2.8 |
| Target/filter | TTS Default | 3.4 |

Informal listening tests with the test data indicate that the intonation model does a good job of capturing the style of speaking in that others who were not familiar with this work, but were familiar with the corpus, could recognize the speaker. Though the speaker style is partly due to the use of natural durations and symbolic prosodic labels used to generate the speech, the speaker is not easily identifiable under those same conditions with the target/filter $F_0$ contour.

## VI. SUMMARY

A new method has been presented for the generation of $F_0$ and energy for speech synthesis, based on a nontraditional statistical approach using a dynamical system model for generating phrase-level contours and a regression tree for predicting a scaling factor for each phrase. Quantitative and perceptual results show that this approach is capable of producing good approximations to observed $F_0$ and log energy contours using lexical information and symbolic prosodic labels. This method of generating $F_0$, used in conjunction with a model for predicting the symbolic prosodic labels from text (e.g., [6], [7]), could produce a text-to-speech synthesizer with a more natural intonation. The model for energy did not improve speech quality, but there are many parameters that could be adjusted since the dependencies were designed primarily with $F_0$ modeling in mind.

Despite the novel statistical approach, the model builds on advantages of other work, i.e., the smoothing function of the filter in the command/filter model [25], [26], the representation of $F_0$ targets [18], [29], and the multilevel representation of $F_0$ [18] including the modeling of segmental effects. Our model has the advantage over these purely rule-based models in that it can be automatically trained. Since the structure is based in linguistic theory, it also covers a broader range of prosodic characteristics than do other automatically-trainable models, e.g., [38]. A potential advantage of the dynamical system model is that it can handle vector processes, allowing for joint modeling of energy and $F_0$. While the joint model did not give significant gains in the experiments reported here, experiments in intonation recognition (with HMM's) suggest that a joint $F_0$/energy model is useful. Finally, the architecture presented here can be easily scaled in complexity by adding or omitting parameter dependencies, given sufficient training data, so the model can be tailored for use on different platforms or to different phonological theories. The main disadvantage of this general approach is that it requires a large amount of labeled data, and hand labeling prosodic markers is very expensive. However, much progress is being made on the problem of automatic labeling (e.g., [39]), and this work is itself useful in improving the accuracy of automatic prosodic labeling using maximum likelihood recognition [40].

While this work represents an important advance in intonation modeling, the primary contribution is the structural framework. There is still much work that can be done to improve on the results presented here. Further exploration of intrasyllable timing and parameter dependencies within the dynamical system model, particularly for energy modeling, will undoubtedly lead to better performance. Pitch range prediction could also be improved with better predictor variables, particularly if a discourse analysis was available as in message-to-speech generation. Finally, it would be interesting to assess the methodology on other speakers or speaking styles, to confirm initial impressions that style is captured well by this model.

## Acknowledgment

The radio news corpus was provided by WBUR, a Boston public radio station. The authors gratefully acknowledge L. Dilley's work in labeling much of the data, and thank N. Campbell and M. Beckman for comments on an earlier version of the manuscript, and AT&T for use of their synthesizer and for providing the capability to test $F_0$ and energy contour generation.

## References

[1] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer*, vol. 82, pp. 737–793, 1987.

[2] R. Collier, "Multi-language intonation synthesis," *J. Phonet.*, vol. 19, pp. 61–73, 1991.

[3] K. Silverman, "On customizing prosody in speech synthesis: Names and addresses as a case in point," in *Proc. ARPA Workshop on Human Language Technology*, 1993, pp. 317–322.

[4] M. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Comput. Speech Lang.*, vol. 6, pp. 175–196, 1992.

[5] M. Ostendorf and N. Veilleux, "A hierarchical stochastic model for automatic prediction of prosodic boundary location," *Computat. Linguist.*, vol. 20, pp. 27–54, 1994.

[6] J. Hirschberg, "Pitch accent in context: Predicting prominence from text," *Artif. Intell.*, vol. 63, pp. 305–340, 1993.

[7] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Comput. Speech Lang.*, vol. 10, pp. 155–185, 1996.

[8] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 431–442, 1993.

[9] K. E. A. Silverman and J. B. Pierrehumbert, "The timing of prenuclear high accents in English," in *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, J. Kingston and M. Beckman, Eds. Cambridge, U.K.: Cambridge Univ. Press, 1990, pp. 72–106.

[10] D. O'Shaughnessy, "Linguistic features in fundamental frequency patterns," *J. Phonet.*, vol. 7, pp. 119–145, 1979.

[11] P. Lieberman *et al.*, "Measures of the sentence intonation of read and spontaneous speech in American English," *J. Acoust. Soc. Amer.*, vol. 77, pp. 649–657, 1985.

[12] M. Liberman and J. Pierrehumbert, "Intonational invariants under changes in pitch range and length," in M. Aronoff and R. T. Oehrle, Eds., *Language Sound Structure*. Cambridge, MA: MIT Press, 1984, pp. 157–233.

[13] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonol. Yearbook*, vol. 3, pp. 255–309, 1986.

[14] J. Hirschberg, "Studies of intonation and discourse," in *Proc. ESCA Workshop on Prosody: Working Papers 41*, 1993, pp. 90–95.

[15] M. Swerts, R. Geluykens, and J. Terken, "Prosodic correlates of discourse units in spontaneous speech," in *Proc. Int. Conf. Spoken Language Processing*, 1992, pp. 421–424.

[16] B. Grosz and J. Hirschberg, "Some intonational characteristics of discourse structure," in *Proc. Int. Conf. Spoken Language Processing*, 1992, pp. 429–432.

[17] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *J. Acoust. Soc. Amer.*, vol. 90, pp. 2956–2970, 1991.

[18] K. Silverman, "The structure and processing of fundamental frequency contours," Ph.D. dissertation, Univ. Cambridge, U.K., 1987.

[19] J. P. H. van Santen and J. Hirschberg, "Segmental effects on timing and height of pitch contours," in *Proc. Int. Conf. Spoken Language Processing*, 1994, vol. 2, pp. 719–722.

[20] M. Beckman and G. Ayers, "Guidelines for ToBI labeling, version 2.0," Manuscript and accompanying speech materials [Obtain by writing to tobi@ling.ohio-state.edu], 1994.

[21] J. F. Pitrelli, M. E. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *Proc. Int. Conf. Spoken Language Processing*, 1994, vol. 1, pp. 123–126.

[22] K. Silverman *et al.*, "ToBI: A standard for labeling English prosody," in *Proc. Int. Conf. Spoken Language Processing*, 1992, pp. 867–870.

[23] J. Pierrehumbert, "Synthesizing intonation," *J. Acoust. Soc. Amer.*, vol. 70, pp. 985–995, 1981.

[24] C. Traber, "$F_0$ generation with a data base of natural $F_0$ patterns and with a neural network," in *Talking Machines: Theories, Models, and Designs*, G. Bailey, C. Benoît, and T. R. Sawallis, Eds. Amsterdam, The Netherlands: Elsevier, 1992, pp. 287–304.

[25] H. Fujisaki and H. Kawai, "Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, 1988, pp. 663–666.

[26] K. Hirose and H. Fujisaki, "Analysis and synthesis of voice fundamental frequency contours of spoken sentences," in *Proc. Int. Conf. Acoust., Speech and Signal Processing*, 1982, pp. 950–953.

[27] D. Hirst, P. Nicolas, and R. Espesser, "Coding the $F_0$ of a continuous text in French: An experimental approach," in *Proc. 12th Int. Congress of Phonetic Sciences*, pp. 234–237, 1991.

[28] V. Aubergé, "Prosody modeling with a dynamic lexicon of intonative forms: Application for text-to-speech synthesis," in *Proc. ESCA Workshop on Prosody: Working Papers 41*, 1993, pp. 62–65.

[29] M. D. Anderson, J. B. Pierrehumbert, and M. Y. Liberman, "Synthesis by rule of English intonation patterns," in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, 1984, pp. 2.8.1–2.8.4.

[30] J. Pierrehumbert, "The phonology and phonetics of English intonation," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, 1980.

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[32] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA J.*, vol. 3, pp. 1445–1450, 1965.

[33] K. Ross, "Modeling of intonation for speech synthesis," Ph.D. dissertation, Boston University, Boston, MA, 1995.

[34] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth, 1984.

[35] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," Tech. Rep. ECS-95-001, ECS Dept., Boston Univ., Boston, MA, 1995.

[36] D. Talkin, "Pitch tracking," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995.

[37] K. Bartkova, P. Haffner, and D. Larreur, "Intensity prediction for speech synthesis in French," in *Proc. ESCA Workshop on Prosody: Working Papers 41*, 1993, pp. 280–283.

[38] M. S. Scordilis and J. N. Gowdy, "Neural network based generation of fundamental frequency contours," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1989, pp. 219–222.

[39] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 469–481, 1994.

[40] M. Ostendorf and K. Ross, "A Multi-level model for recognition of intonation labels," in *Computing Prosody*, Y. Sagisaka, N. Campbell and N. Higuchi, Eds. New York: Springer-Verlag, 1997, pp. 291–308.

**Kenneth N. Ross** (M'82) received the B.S. degree in electrical engineering from Case Western Reserve University, Cleveland, OH, in 1984, the M.S. degree in electrical engineering from Polytechnic University (formerly Brooklyn Polytechnic), Brooklyn, NY, in 1986, and the Ph.D. degree in electrical engineering from Boston University, Boston, MA, in 1995. While at Boston University, he worked on statistical modeling for speech processing applications, with a specific goal of developing computational models of prosody. The primary focus was the modeling of prosody for text-to-speech synthesis applications, but these models were also applied to the recognition of abstract prosodic labels.

From 1995 to 1997, he was a Research Scientist at BBN Laboratories, Cambridge, MA, on large vocabulary conversational speech recognition in both English and Spanish. He has been at ALPHATECH, Inc., since October 1997, where his main research interest has been in information fusion.

**Mari Ostendorf** (M'85–SM'97) received the B.S., M.S., and Ph.D. degrees in 1980, 1981, and 1985, respectively, all in electrical engineering, from Stanford University, Stanford, CA.

In 1985, she joined the Speech Signal Processing Group at BBN Laboratories, Cambridge, MA, where she worked on low-rate coding and acoustic modeling for continuous speech recognition. She joined the Department of Electrical and Computer Engineering, Boston University, Boston, MA, in 1987, where she is currently an Associate Professor. Her research interests include data compression and statistical pattern recognition, particularly in speech processing applications. Her recent work involves investigation of segment-based and higher order models for continuous speech recognition, and stochastic models of prosody for both recognition and synthesis.

Dr. Ostendorf has served on the Speech Processing and DSP Education Committees of the IEEE Signal Processing Society, and is a member of Sigma Xi.