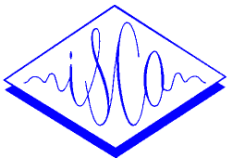

The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26–29, 1998

DESCRIPTION OF THE BELL LABS INTONATION SYSTEM

Jan P. H. van Santen, Bernd Möbius, Jennifer J. Venditti and Chilin Shih

Bell Labs – Lucent Technologies, Murray Hill, NJ, USA.

ABSTRACT

This paper describes the approach to intonation modeling currently used in the Bell Labs Multi-lingual Text-to-Speech System for English, French, German, Italian, Spanish, Russian, Romanian, and Japanese.

1. INTRODUCTION

The intonation component of the Bell Labs Multi-lingual Text to Speech System [11] plays the customary role of computing a fundamental frequency contour from phonological representations consisting of phoneme labels, and symbols for phrasing and accenting. Two aspects of our approach are unique.

First, we model in considerable detail the effects of segments and their durations on the time course of the fundamental frequency contour. The basic idea here is that local pitch excursions associated with pitch accents ("accent curves") are tied to the accented syllable in a complicated, yet tight manner. Two assumptions are central. First, following Möbius [4], the phonological unit of a pitch accent is not the accented syllable, but the accent group, defined as an accented syllable followed by zero or more unaccented syllables. Second, the time course of an accent curve depends on the entire segmental and temporal structure of the accent group, not only on the properties of the accented syllable. While complicated, this dependency is deeply regular, and exhibits the property that, ceteris paribus, the pitch peak (as measured from the start of the accented syllable) is shifted rightward as any part of the accent group is lengthened. As we shall see, this fundamental regularity can be captured to a first order of approximation by a simple linear *alignment model*. A key perceptual advantage of this detailed alignment model is that – in particular under extreme circumstances such as a very long "wow!", early nuclear pitch accents followed by a large number of unaccented syllables, or very strong pitch excursions – intonation contours remain properly aligned with the segment and syllable boundaries.

Second, building on the well-known superpositional model by Fujisaki [2, 1], we considerably broaden the concept of superposition. In the Fujisaki/Ohman model, an F_o contour is generated by addition (in the log domain) of two different types of curves: accent curves and phrase curves. The former are generated by smoothing rectangular accent commands, and the latter by smoothing impulses at the start and end of the phrase. Both types of smoothing are performed by filters having specific mathematical forms. We alter this model as follows. First, as mentioned above, we explicitly tie accent curves to accent groups. Second, we include segmental perturbation curves to describe the very large, but short-lived effects one observes during the initial 50-100 ms of post-obstruental vowels. Third, we loosen the restrictions on the shape of phrase curves, which enables us to adequately model descending curves (most languages) as well rise-fall patterns (Japanese, Russian); in the Fujisaki model, phrase curves do not allow for a substantial variety of shapes. And, fourth, of course, accent curves are generated via a linear alignment model, not via a smoothed rectangular accent command.

In this paper we review some of our work on alignment in English. We then show – at a somewhat anecdotal level – how our model can be applied to certain contours in Japanese. A description of the current implementation in the Bell Labs Multi-lingual Text to Speech System is presented.

2. Alignment in American English.

The starting point of the research underlying the current Bell Labs intonation system is a series of studies on alignment in American English [10], where the following key findings were obtained. In these studies, we used speech recorded from a female speaker who produced carrier phrase utterances in which one or two words were systematically varied. The non-varying parts of the utterances contained no pitch accents. One study focuses on utterance-final monosyllabic accent groups, produced with a single "high" pitch accent, a low phrase accent, and a low boundary tone (Pierrehumbert label H*LL% [5]. Other studies also used for polysyllabic accent groups, continuation contours (H*LH%), and yes/no contours (L*HH%).

2.1. Results on simple peak placement rules

Initially, we measured peak location (from accented syllables start) for the H*LL%, and found that peak location varied systematically as a result of the phonetic class of the onset (voiceless, voiced obstruent, sonorant) and of the coda (voiceless, voiced obstruent, sonorant, polysyllabic). For example, peak location is systematically later in sonorant-final accent groups than in obstruent-final accent groups (*pin* vs. *pit*), and later in obstruent-initial accent groups than in sonorant-initial accent groups (*bet* vs. *yet*). Such effects persisted when we measured peak location from vowel start instead of syllable start, and when we normalized peak location by division by syllable or rhyme duration. Apparently, peaks are located at neither a fixed millisecond amount nor a fixed fraction of the accent group, even if we restrict attention to monosyllabic accent groups. Polysyllabicity had particularly strong effects on peak placement when measured in terms of relative location in the accented syllable: peaks occur much later in the initial accented syllable (91% on average, and often located in the second syllable) compared to monosyllabic accent groups (35%); relative to the entire accent group, peaks occur significantly earlier in polysyllabic accent groups (35%) than in monosyllabic accent groups (45%).

While these results showed that certain simple peak placement

rules do not work, they did not immediately suggest rules that might work.

2.2. Linear alignment model for peak placement

Simple peak placement rules can all be viewed as special cases of a *linear alignment model*, which we explain next. Suppose we split the temporal interval corresponding to an accent group into multiple sub-intervals, such as those that correspond to the accented syllable onset, the accented syllable rhyme, and the remainder of the interval spanned by the accent group. Call the durations of these three respective intervals $D_1(a)$, $D_2(a)$, and $D_3(a)$, where a refer to a particular rendition of the accent group in question (e.g., a rendition of the accent group corresponding to the word "temporal"). Thus, $D_1(a)$ is the duration of the initial /t/, $D_2(a)$ is the duration of the /em/ rhyme, and $D_3(a)$ is the duration of /poral/.

We now propose that peak time depends linearly on these three durations. Formally,

$$T_{peak}(a) = \sum_j \alpha_{\mathbf{S},j} \times D_j(a) + \mu_{\mathbf{S}}. \quad (1)$$

Here, $T_{peak}(a)$ is peak location, j refers to the j-th "part" of the accent group, $D_j(a)$ is the corresponding duration, and $\alpha_{\mathbf{S},j}$ its weight. We refer by *accent group segmental structure*, **S**, to the sequence of classes corresponding to the segments in the accent group (see below). The intercept $\mu_{\mathbf{S}}$ depends only on segmental structure (**S**), and is routinely included in linear modeling.

Note that we include any non-syllable-initial sonorants in the accented syllable rhyme. For codas in monosyllabic accent groups, we include only the sonorants in the rhyme. Thus, the rhyme is *lan* in *blank*, /i/ in *seat*, /yu/ in *muse*, /in/ in *seen*, and /o/ in *off*; but in the word *offset*, the rhyme consists of /of/. We distinguish between four types of segmental structure: monosyllabic (coda: sonorant, voiceless, voiced obstruent) vs. polysyllabic. In other words, we do not distinguish between different onset classes; this was based on the finding of no differences in weights between different onset classes.

To illustrate, *blank*, *seat*, and *off* have the same structure (monosyllabic, voiceless coda), while *muse* and *seen* are examples of the other two monosyllabic types; the final two syllables of *syllabic* have the polysyllabic type.

We now explain how this model generalizes various simple peak placement rules. The hypothesis that peak placement is solely determined by overall accent group duration corresponds to the statement that

$$\alpha_{\mathbf{S},j} = \alpha, for\ all\ \mathbf{S}\ and\ j. \quad (2)$$

Another rule is that peaks are placed a fixed fraction (F) into the rhyme. For this, we let

$$\begin{aligned} \alpha_{\mathbf{S},1} &= 1 \\ \alpha_{\mathbf{S},2} &= F \\ \mu_S &= 0.0 \end{aligned} \quad (3)$$

The rule that the peak is placed a fixed ms amount (M) into the vowel (as IPO approach [8]) is given by

$$\begin{aligned} \alpha_{\mathbf{S},1} &= 1 \\ \alpha_{\mathbf{S},j} &= 0 \\ \mu_S &= M \end{aligned} \quad (4)$$

Of course, the fact that the linear alignment model generalizes various common simple rules, while of conceptual interest, is no reason for its acceptance. We present two arguments in favor of this model. One is theoretical, the other empirical

Sub-interval duration directional invariance The model is an instantiation of a much broader concept, namely that of *sub-interval duration directional invariance principle.* According to this principle, for any two accent groups a and b that have the same segmental structure:

$$If\ D_j(a) \geq D_j(b)\ for\ all\ j\ then\ T_{peak}(a) \geq T_{peak}(b). \quad (5)$$

Our alignment model is a special case, because when

$$D_j(a) \geq D_j(b)\ for\ all\ j$$

then, because all α parameters are non-negative:

$$\sum_j \alpha_{\mathbf{S},j} D_j(a) \geq \sum_j \alpha_{\mathbf{S},j} D_j(b)$$

and hence, by definition of our model (Equation 6):

$$T_{peak}(a) \geq T_{peak}(b).$$

The principle simply states that *stretching any "part" of an accent group has the effect of moving the peak to the right*, regardless of whether the stretching is caused by speaking rate changes, contextual effects on the constituent segments (e.g., degree of emphasis), or intrinsic duration differences between otherwise equivalent segments (e.g., /s/ and /p/ are both voiceless and hence equivalent, but /s/ is significantly longer than /p/.)

We claim that this principle is not unreasonable for the contours considered. In fact, we find it difficult to imagine thing being different. Note that the principle includes cases where we measure peak location (or, more broadly, *anchor point* location; see below) from the end of the accent group.

Of course, the assumption of linearity goes beyond the general principle, and is unlikely to be correct. For example, it predicts that a change in the duration of the remaining unstressed syllables from 125 to 150 ms has exactly the same effect on peak location as a change from 525 to 550 ms; in both cases, the size of the effect is ($\alpha_{\mathbf{S},1} \times 25$). We strongly suspect that the second effects will be smaller. This can be fixed by applying a negatively accelerated transformation to the subdurations (e.g., taking $D_j(a)^{0.5}$.) The broad acceptance of linear models in many areas of research rests on the fact that the linear model is an excellent first-order approximation in situations where factors (here: subdurations) have directionally invariant effects on the variable of interest (here: peak location).

Fitting the linear alignment model The parameters of the model (α, μ) can be estimated using standard linear regression methods, because the quantities D and T_{peak} are directly measurable. Consequently, the model in fact provides a convenient framework for testing the simple peak placement rules.

Results showed the following. First, the overall fit is quite good. The predicted-observed correlation of 0.91 (r^2 = 83%) for peak location explains more than 2.3 times the variance explained by overall accent group duration, where the correlation was 0.59 (r^2 = 35%).

Second, the weights $\alpha_{\mathbf{S},j}$ varied strongly as a function of part location (j = *onset, rhyme, remainder*), with the effects of the onset being the strongest and the effects of the remainder being the weakest.

Third, setting the intercept $\mu_{\mathbf{S}}$ to zero did not affect the fit (it reduced r^2 from 83% to 81%), suggesting that the accented syllable start plays a pivotal role in alignment. This analysis also contradicts the rule that the peak is placed a fixed ms amount into the vowel (Eq. 4).

Fourth, the values of the $\alpha_{\mathbf{S},j}$ parameters depended on segmental structure. Specifically, the values of the onset weights, $\alpha_{\mathbf{S},1}$, were smaller for sonorant *codas* than for non-sonorant codas; however, the onset weights were the same for all onset types, and ranged from 0.60 for sonorant codas to values in the 0.85-1.0 range for the other coda types (approximate values are given because the values dependent somewhat on details of the regression algorithm). The fact that $\alpha_{\mathbf{S},1}$ is less than 1.0 violates the rule that peak are placed a fixed fraction (F) into the rhyme (Eq. 3). A stronger violation of the same rule is, of course, that the peak is much later (measured as a fraction of the accented syllable) in polysyllabic than in monosyllabic accent groups.

2.3. Linear alignment model for anchor points

The peak is only one point on an accent curve, and it is not clear whether it is the most important point – perhaps it is the start of the rise, or the point where the rise is steepest. In the tone sequence tradition following Pierrehumbert (1980), tone targets are elements of the phonological description, whereas the transitions between the targets are described by phonetic realization rules. Taking the opposite view, the IPO approach ('t Hart et al., 1990) assigns phonological status to the transitions, viz. the pitch accent movement types, themselves. In the model proposed here, both pitch movements and specific points characterizing the shape of the movements matter. However, no particular status is reserved for peaks; our "targets" are non-linear pitch curves, which are often either bidirectional (H*LL% contour) or tridirectional (continuation contour). One way to capture the entire curve is by sampling many points on that curve ("anchor points"), and model timing of these points in the same way as peak location.

We have experimented with various methods for defining such points on an accent curve. For example, one can compute the first or second derivative of the accent curve, and define as anchor points locations where these derivatives cross zero or reach other special values. However, derivatives are not particularly well-behaved in the case of F_0 curves due to small local variations in periodicity. Among the methods that we used, the following proved to be the simplest and at the same time statistically most robust. We subtract a locally straight "phrase curve" from the observed F_0 curve around the area where the accent curve is located, and then consider the residual curve as an estimate of the accent curve (*estimated accent curve*). For the H*H%L curves, the locally straight "phrase curve" is computed simply by connecting the last sonorant frame preceding the accent group with the final sonorant frame of the accent group. We then sample the estimated accent curve at locations corresponding to a range of percentages between 0% and 100% (e.g., 5%, 10%, 25%, ..., 75%, 90%, 95%, 100%) of maximal height. Thus, the 100% point is the peak location, and the 50% pre-peak point is the time point where the estimated accent curve is half of maximal height. We call these time points *anchor points*.

The model in Equation (1) can be applied to any anchor point by replacing the *peak* subscript by i (for the i-th anchor point) and adding i as a subscript to the parameters α and μ:

$$T_i(a) = \sum_j \alpha_{i,\mathbf{S},j} \times D_j(a) + \mu_{i,\mathbf{S}}\,. \qquad (6)$$

We call the ensemble of regression weights ($\alpha_{i,\mathbf{S},j}$), for a fixed segmental structure **S**, the *alignment parameter matrix* (*APM*), and Equation 6 the *alignment model*. It could be said that an *APM characterizes for a given pitch accent type how accent curves are aligned with accent groups.*

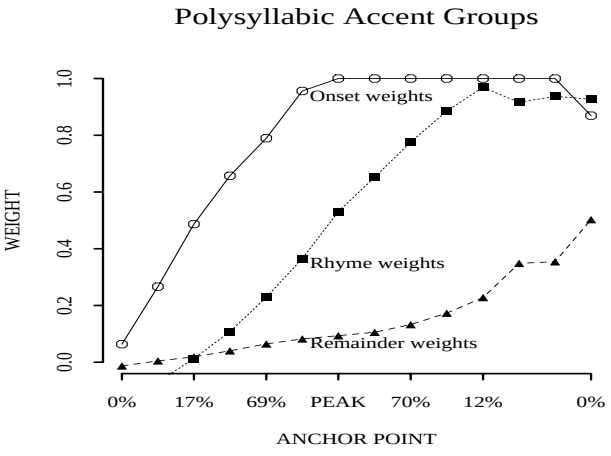

Figure 1: Regression weights as a function of anchor point, for each of the three sub-intervals of the accent group. H*LL% accent type. Solid curve: onset; dotted curve: rhyme, dashed curve: remainder.

Figure 1 shows the values of the alignment parameters for polysyllabic phrase-final accent groups (H*LL%). We note the following. First, the weights for the onset exceed the weights for the rhyme, and the latter exceed the weights for the remainder of the accent group. In other words, lengthening the onset duration of the stressed syllable by a fixed ms amount has a larger effect on

any anchor point than lengthening the duration of the unstressed syllables by the same ms amount. Second, the curves are monotonically increasing. They initially diverge, and then converge. Early anchor points mostly depend on onset duration and hardly on the durations of the rhyme and the remainder, but late anchor points depend more evenly on all three subsequence durations. A key point is that these alignment curves are well-behaved, and without a doubt can be captured by a few meta-parameters, e.g., two straight line segments per curve.

2.4. Alignment Parameters and Time Warping

The alignment model was described as a way to predict the locations of certain points on the F_0 curve from the durations of judiciously selected parts of an accent group, via multiple regression. However, we can re-cast the model in terms of *time warping of a common template.*

Assume that we normalize the accent curve to the 0-1 interval. According to the model, this curve is given by:

$$\hat{F}_0(t) \;=\; P[i(t)]. \tag{7}$$

Here, t represents time; $i = i(t)$ is the index onto which location t is mapped, using an appropriate interpolation scheme. $P(i)$ is the percentage corresponding to the i-th anchor point. In Panel (f), t corresponds to the horizontal axis and $i(t)$ to the vertical axis.

We further clarify the relationship between alignment parameters and time-warping by spelling out the steps involved in the computation of the normalized accent curve for a rendition a of a given accent group:

Step 1: Measure the durations of all subintervals D_j of a.

Step 2: For each anchor point i, compute predicted anchor point location T_i using Equation 6.

Step 3: For each time point t, find i such that t is located between T_i and T_{i+1} .

Step 4: Retrieve values P_i and P_{i+1}, and obtain a value for t by interpolation.

In summary, the normalized accent curve can be viewed as *time warped version of a common template*. The time warps for a given accent curve class vary from one utterance to the next, but they belong to the same *family* in that they are all are produced by Equation (6).

3. Other Segmental Effects in American English

3.1. Perturbation curves

As stated in the Introduction, we are concerned with the short-lived, but strong, effects on F_o curves during the initial part of post-obstruental vowels, for two reasons. First, because of their strength, these effects can hinder analysis by creating spurious local peaks or obscuring true peaks. Second, we are not convinced that these effects can be ignored during synthesis. Perhaps in the early days of single-pulse LPC synthesis, these effects were not audible. But in the presence of constant improvements of voice quality, we are much less confident.

Perturbation curves are associated with initial parts of sonorants following a transition from an obstruent. We measured these effects, by contrasting vowels preceded by sonorants, voiced obstruents, and unvoiced obstruents in syllables that were not accented and were not preceded in the phrase by any accented syllables [10]. These curves are described by a rapid decay from values of about 30 Hz to 0 in 100 ms, and are added in the logarithmic domain to the other curves.

3.2. Intrinsic pitch

It is well-known (e.g., [7]) that there are also effects of segmental identity on the time course of F_o during the segment itself (vs. on F_o during the following segment, as in segmental perturbation). Effects include vowel height, and nasality. Interestingly, we have found large individual differences in the effects of nasality, with some speakers showing no effect at all (such as the speaker used for all studies discussed so far), and other speakers showing 5-10 Hz depressions that are perfectly aligned with the start and end of nasal regions.

All these effects can be modeled by adding or subtracting segment-specific constants in the log domain. One complication we noted is a possible interaction with accent status: In our key speaker, we found intrinsic pitch effects only in accented syllables.

4. Current implementation.

The implementation of this model in the Bell Labs TTS system is relatively straightforward [11].

Accent groups and minor phrases, and segmental durations, are pre-computed by preceding modules in the TTS system. Based on this information, the duration module computes the phrase curves, accent curves, perturbation curves, and intrinsic F_o parameters, and combines these via the superposition principle.

What we have not discussed yet in the preceding sections is the computation of phrase curves and the determination of accent curve height or amplitude.

4.1. Phrase curves

For English, we found that phrase curves could be modeled as two-part curves obtained by non-linear interpolation between three points, viz. the start of the phrase, the start of the last accent group in the phrase, and the end of the phrase.

The phrase curve model includes as special cases the standard linear declination line, and curves that are quite close to the phrase curve in Fujisaki's model. Moreover, some of the problems with the Fujisaki model, especially its apparent inability to model certain contour shapes observed in English (see discussion by [3], p. 30), can be attributed to too strong constraints on the shape of commands and contours. We prefer to be open to the possibility that phrase curves exhibit considerable and meaningful variabil-

ity. Phrase curve parameters are controlled by sentence mode and locational factors, such as sentence location in the paragraph.

4.2. Accent curve height

In our model, accent curve height is determined via a multiplicative model by multiple factors, including position (in the minor phrase, the minor phrase in major phrase, etc.), factors predictive of prominence, and intrinsic pitch. The multiplicative model is often used in segmental duration modeling. It makes the important – and not necessarily accurate – assumption of directional invariance [9]: holding all factors but one constant, the effects of the varying factor always have the same direction. This may often be true in segmental duration; e.g., when two occurrences of the same vowel involve identical contexts, except for syllabic stress, the stressed occurrence is likely to be longer. However, one has to be very careful in which factors one selects and how one defines them. For example, if one were to use as factors the parts-of-speech of the word in question and its left and right neighbors, such directional invariance is extremely unlikely to occur.

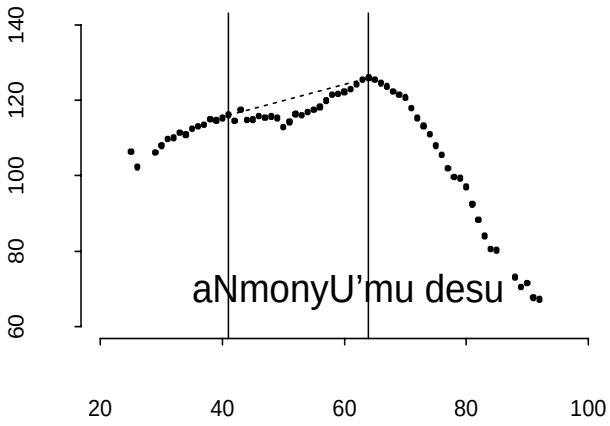

Figure 2: Fundamental frequency contour for the word "ammonium", in Japanese. The horizontal axis is time (in ms), the vertical frequency (in Hz). Vertical lines show the two peaks, and the dashed line connecting these peaks highlights the locally concave shape.

5. Exploration of Japanese Intonation

The approach we have taken to Japanese differs in two key respects from the languages discussed above, having to do with important differences in prosodic organization. Specifically, we use the concept of *UA group* to refer to sequences of words such that the first n are unaccented ($n \geq 0$) and either the last word is accented or the last word is followed by a phrase boundary. Thus, a UA group can be a single unaccented or accented word followed by a phrase boundary, three unaccented words followed by an accented word, or a single accented word. The UA group is similar to Fujisaki's 'accent phrases' [2], Accent groups are defined in the usual way, replacing syllables by morae.

In our approach, we construct a *UA curve* for each UA group, and an accent curve for each accent group. UA curves are computed in a way that differs sharply from the way minor phrase curves are computed in other languages. First, rather than descending throughout, they start with an ascending portion followed by a descending portion. Second, their shape is affected by far more factors than the minor phrase curves in the other languages. These factors include accent status (i.e., does the UA group consist solely of unaccented words, in which case it must be terminated by a phrase boundary), phonological length of the initial syllable, and the location of the accented mora in the UA group.

In addition to the accent curves and UA curves, we also use near-linear minor phrase curves to model global declination trends.

A key motivating factor for our approach here is the observation that the shape of fundamental frequency contours in accented UA groups often exhibits two local peaks (see Figure 2. The first is often reached around the end of the second mora, and the second peak occurs at the end of the accented mora and is followed by a steep decline; often, the second peak is not formally a local maximum, but consists of a gentle decline followed by a sharp decline. In either case, however, the curve segment between the initial peak and the point where the steep decline starts (at the end of the accented mora) *is not a straight line*, but is below the straight line connecting these two points – it is *concave*. More importantly, we noted that different renditions of the same word often varied in the curvature of this curve segment, with at times the second peak being a clear local maximum (as in the Figure) and at other times just an inclination point where a gentle decline becomes steep. Perceptually, these differences correspond to different degrees of prominence. The obvious way to model this shape is by positing two underlying curves whose heights vary quasi-independently – a UA curve and an accent curve. This interpretation, of course, differs from the Fujisaki model (which models accented regions by single smoothed rectangular accent commands, which cannot be locally concave; it hence often uses two such commands, which is not elegant).

It also differs from the Beckman/Pierrehumbert approach [6] which proposes a linear interpolation between the early phrasal H- (scaled lower in the range) and the following accent H* (scaled at the top of the local range). In our account, the height of the UA curve maximum and that of the accent peak (H*) are not kept in constant relation, but rather can be varied independently, producing a continuum of prominence relations between the peaks. This allows us to use a single underlying structure to account for cases which would need to be analyzed as either one or two accentual phrases in both the Fujisaki and Beckman/Pierrehumbert accounts.

6. Conclusions.

We have described some of the empirical results underlying the approach to intonation control in the Bell Labs TTS System. The same intonation module is used for German, French, Italian, Spanish, Romanian, and Russian. The key adjustments that had to be made for these languages concerned transformations of the alignment parameter matrices. These transformations independently vary the steepness of rises and falls, and involve matrix operations using a small number of meta-parameters. These ad-

justments were based on informal listening experiments.

It should be understood that we consider the current implementation as a promising start, but with many research questions left open to be answered. These questions range from phonological issues (e.g., is it indeed a viable idea to link phonological accent classes to alignment parameter matrices), to detailed phonetic issues (e.g., intrinsic pitch). Our hope is that the current model can serve as a framework for a style of intonation research which more closely integrates phonology and phonetics than is often currently the case.

7. REFERENCES

1. Fujisaki, H. Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech*, P. F. MacNeilage, Ed. Springer, New York, 1983, pp. 39–55.
2. Fujisaki, H., and Sudo, H. Synthesis by rule of prosodic features of connected japanese. In *International Congress on Acoustics*. (Budapest, Hungary, 1971), pp. 133–136.
3. Ladd, D. *Intonational phonology*. Cambridge University Press, Cambridge, UK, 1996.
4. Möbius, B., Pätzold, M., and Hess, W. Analysis and synthesis of German F0 contours by means of Fujisaki's model. *Speech Communication 13* 1993.
5. Pierrehumbert, J. *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1980.
6. Pierrehumbert, J., and Beckman, M. *Japanese Tone Structure*. The MIT Press, Cambridge, Massachussetts, 1988.
7. Silverman, K. *The Structure and Processing of Fundamental Frequency Contours*. PhD thesis, Cambridge University, Cambridge UK, 1987.
8. 't Hart, J., Collier, R., and Cohen, A. *A Perceptual Study of Intonation*. Cambridge University Press, Cambridge UK, 1990.
9. van Santen, J. Prosodic modeling in text-to-speech synthesis. In *Proceedings of Eurospeech-97* (Rhodes, September 1997).
10. van Santen, J., and Hirschberg, J. Segmental effects on timing and height of pitch contours. In *Proceedings ICSLP '94* (1994), pp. 719–722.
11. van Santen, J., Shih, C., and Möbius, B. Intonation. In *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, R. Sproat, Ed. Kluwer, Boston, MA, 1997, ch. 6, pp. 141–189.