A Model of Fundamental Frequency Contour Alignment

Jan P. H. van Santen and Bernd Möbius

Abstract

Segmental factors can cause large temporal changes in local pitch excursions associated with accented syllables ("accent curves"), but these changes are often not phonologically or perceptually significant. Yet, other factors can cause temporal changes that are smaller but nevertheless significant. We propose a model according to which accent curves are (phonologically, perceptually) equivalent when they are generated from a common accent curve template using a common set of alignment rules. These alignment rules specify where the points making up the template are to be placed on the time axis, based on the durations of the segments in the syllable sequence with which the accent curve is associated. The model is shown to provide a detailed and accurate account of alignment of accent curves over a wide range of segmental configurations. This pitch accent model is embedded in a superpositional framework, in which accent curves, segmental perturbation curves, and phrase curves are combined to account for complicated surface F_0 curves. Height of accent curves is determined via a multiplicative model from factors related to prominence and position in the phrase.

1 Introduction

Local pitch contours belonging to the same perceptual or phonological class vary significantly as a result of the structure (i.e., the segments and their durations) of the syllables they are associated with. For example, in nuclear rise-fall pitch accents in declaratives, peak location (measured from stressed syllable start) can vary systematically between 150 and 300 ms as a function of the durations of the associated segments (van Santen and Hirschberg 1994). Yet, there are temporal changes in local pitch contours that are phonologically significant even though their magnitudes do not appear to be larger than changes due to segmental effects (e.g., (Kohler 1990; d' Imperio and House 1997)).

This paper starts out by addressing the following question: *What is invariant about the alignment of pitch accent contours belonging to the same class?* (We losely define a *pitch*

accent contour as a local pitch excursion that corresponds to an accented syllable.) We propose a model according to which pitch accent curves in the same class are generated from a common template using a common set of *alignment parameters.* These alignment parameters specify how the time course of these curves depends on the durations of the segment sequence with which a pitch accent is associated. The paper presents data leading up to this model, and defines precisely what alignment parameters are. We then proceed to embed this model in a more complete intonation model, that in key respects is similar to – but also different from – the superpositional model by Fujisaki (Fujisaki 1983). Our model, besides serving the practical purpose of being used for most languages in the Bell Labs text-to-speech system, is also of conceptual interest, because in a natural way it leads to a broader, yet reasonably precise, definition of the superposition concept. In our experience, discussions of tone sequence vs. superpositional approaches to intonation (e.g., (Ladd 1996)) often suffer from too narrow definition of the superposition concept.

2 Accent Curve Alignment

To keep this section as empirical and theory-free as possible, the word "accent curve" is used very loosely in the sense of a local pitch excursion that corresponds to an accented syllable, not in the specific sense of the Fujisaki model (Fujisaki 1983). In what follows, the term "accent group" (or "stress group") refers to a sequence of syllables of which only the first is accented. "Accent group structure" refers to the segments in an accent group ("segmental structure") with associated durations. Thus, renditions of the same accent group almost always have different structures because their timing is unlikely to be identical, but by definition they have the same segmental structure.

Our data base is an extension of the speech corpus described in a previous paper (van Santen and Hirschberg 1994), and consists of 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. The earlier study focused 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% (Pierrehumbert 1980); Figure 1, left panel). The current data base also includes H*LL% contours for polysyllabic accent groups, continuation contours (H*LH%), and yes/no contours (L*H%). Continuation contours consist of a dual motion in which an early peak is followed by a valley and a final rise (Figure 1, center panel). Yes/No contours (Figure 1, right panel) consist of a declining curve for the pre-accent region (not shown), an accelerated decrease starting at the onset of the accented syllable, and then a steep increase in the nucleus. Unless stated otherwise, results are reported for H*LL% contours.

2.1 Effects of accent group duration

As point of departure we take the most obvious analysis: *Measure alignment of H*LL% accent curves in terms of peak location*, and *assume that accent group structure can be captured simply by total duration.* There is indeed a statistically significant correlation

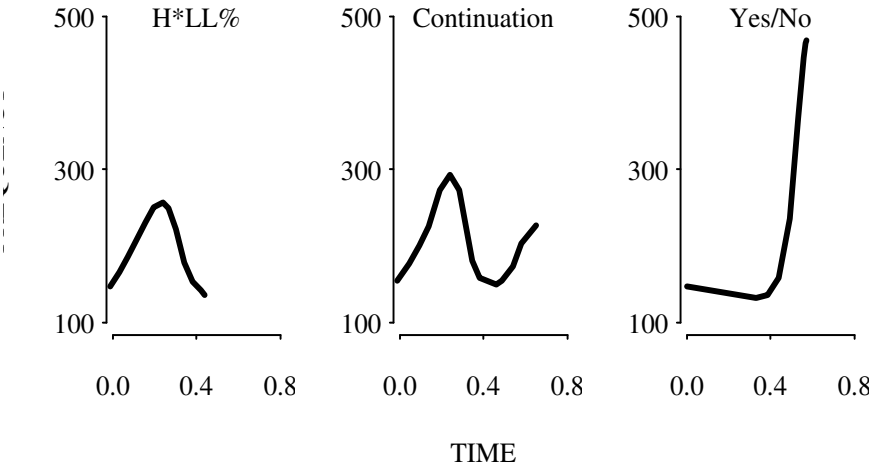

Figure 1: **Averages of Declarative, Continuation, and Yes/No contours.**

between peak location and total duration, showing that peaks are not placed either a fixed or random millisecond amount into the stressed syllable. But the correlation is weak (0.57). We could stop here, and declare that accent curve timing is only loosely coupled to accent group structure. Or, as we do next, we can measure whether timing depends on aspects of accent group structure other than total duration.

2.2 Effects of segmental structure

In (van Santen and Hirschberg 1994) it was shown that peak location strongly depends on segmental structure. For monosyllabic accent groups, peak location (measured from accented syllable start) is systematically later in sonorant-final accent groups than in obstruent-final accent groups (*pin* vs. *pit*), and later in voiced 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.

In our new data, we found that polysyllabic accent groups again behave differently. For example, peaks occur much later in the initial accented syllable (91% of syllable duration on average, and often located in the second syllable) compared to monosyllabic accent groups (35% of syllable duration). Relative to the entire accent group, peaks occur significantly earlier in polysyllabic accent groups (35% of accent group duration) than in monosyllabic accent groups (45% of accent group duration).

2.3 Effects of accent group "sub-durations"

While these data undermine certain peak placement rules used in text-to-speech synthesis (e.g., the rule that peaks are placed a fixed percentage into the accented syllable), they do

not unambiguously disqualify the overall accent group duration hypothesis: overall duration tends to be longer for *pin* than for *pit* (because of the lengthening effect of postvocalic consonant voicing), and longer for "bet" than for "yet" (because /b/ is longer than /y/). In addition, the hypothesis does not require that peaks are located at a fixed fraction into the accented syllable or its rhyme; it only requires that peak locations in accent groups of equal length are the same.

A better test concerns the prediction that changes in peak location do not depend on which *part* of an accent group is lengthened. To illustrate, consider two monosyllabic accent groups that have the same overall duration of 400 ms, but the first (*stick*) has a relatively long onset of 170 ms and a vowel of 180 ms, while the second (*woke*) has a short onset of 60 ms and a longer vowel of 290 ms. In both cases, the duration of the final /k/ is the same (50 ms). If total accent group duration is the sole variable that matters, than both accent curves should be the same. But if alignment depends on more detailed aspects of the temporal structure, than the curves could differ. Our analysis presented next will show that, in fact, the peak in *stick* lies 83 ms to the right of the peak in *woke* (198 vs. 115 ms from the syllable start).

We measure the effects on peak placement of different parts of the accent group by defining the parts, predicting peak location by a weighted combination (multiple regression analysis) of the durations of these parts ("sub-durations"), and inspecting the values of the weights:

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

Here, a is a rendition of an accent group with segmental structure $\mathbf{S}$, $T_{peak}(a)$ 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 use three "parts": accented syllable onset, accented syllable rhyme, and remaining unstressed syllables (polysyllabic accent groups only). 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. This, *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.

This unusual partition of the syllable is based on analyses where we applied Equation 1 for much narrower classes of phonemes, while varying which parts of the syllable were included in the onset or rhyme. We found that the proposed partition provided the most parsimoneous fit.

Equation 1 is strong in that it assumes linearity [referring to the $\sum$ sign]. Thus, it predicts that a change in onset duration from 50 to 75 ms has exactly the same effect on peak location as a change from 125 to 150 ms; in both cases, the size of the effect is ($\alpha_{\mathbf{S},1} \times 25$). Otherwise, it is quite general and subsumes many models and rules proposed in the literature. For example, the hypothesis that peak placement is solely determined by

overall accent group duration corresponds to the statement that

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

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

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

The rule that the peak is placed a fixed ms amount (M) into the vowel (as IPO approach ('t Hart *et al.* 1990)) is given by

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

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 these 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, for all three contour classes, 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. This violates the the hypothesis that peak placement is solely determined by overall accent group duration (Eq. 2), which requires that the weights should be the same.

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, as was reported in Subsection 2.2.

To apply this model to the hypothetical "stick" and "woke" examples, using $\alpha_1 = 0.95$ and $\alpha_2 = 0.2$, we find that:

$$\begin{aligned} T_{peak}(stick) &= 0.95 \times 0.17 + 0.2 \times 0.18 = 0.1975 \\ T_{peak}(woke) &= 0.95 \times 0.06 + 0.2 \times 0.29 = 0.115 \end{aligned}$$

In other words, in *stick* the peak is predicted to occur 197 ms after the start of the accented syllable and in *woke* 115 ms into the syllable. This difference is due to the effects of onset duration (0.95) to be much larger than the effects of the rhyme duration (0.2).

2.3.1 Importance of accented syllable start

The key reason for measuring time starting at accented syllable onset is that in these data the pitch movement appears to start at this time point. When we re-did the analysis with a different starting point, the vowel start, results were far less clear cut. The lower prediction accuracy is in part due to the fact that we lose a free parameter ($\alpha_{\mathbf{S},1}$) when we remove onset duration from the equation. However, the more important reason for the poorer fit is that, even when we hold rhyme duration constant, peak location is not a fixed ms amount after vowel start; in fact, as we reported above for sonorant codas, each 1.0 ms lengthening of the onset causes only a 0.6 ms rightward shift in peak location. This effect cannot be captured by a model that ignores onset duration, such as a model assuming that we should measure time from vowel start.

Independent evidence for the assumption that the pitch movement starts at syllable onset was provided by our finding that the intercept μ was statistically zero.

The importance of the start of the accented syllable in rise-fall curves confirms earlier results by Caspers, who found over a wide range of conditions (in terms of speaking rates, intrinsic duration of the accented vowel, and presence versus absence of nearby pitch movements) that the start of the rise coincides with the start of the accented syllable (Caspers 1994). The start of the rise was not strongly tied to other segment boundaries, and the end of the rise (which in most cases is the same as the peak) was not tied to any segment boundary. This is consistent with our model, because for polysyllabic accent groups peak location is given by:

$$\alpha_{\mathbf{S},1} \times D_{onset} + \alpha_{\mathbf{S},2} \times D_{rhyme} + \alpha_{\mathbf{S},3} \times D_{remainder} + \mu_{\mathbf{S}} \quad (5)$$

the end of the rhyme by $D_{onset} + D_{rhyme}$, and the end of the onset by D_{onset}. Given our estimates of the $\alpha_{\mathbf{S},j}$ and $\mu_{\mathbf{S}}$ parameters, peak location cannot coincide with, or be at a fixed distance from, either of the latter two boundaries.

In a study of Greek pre-nuclear accents, the onset of the rise of rise-fall accents (with peak location in the post-accentual vowel) was also found to coincide with accented syllable start (Arvaniti *et al.* 1998).

2.4 Anchor points

2.4.1 Estimation of 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 typically 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*LH% 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*. In Section 2.4.3 we will discuss how phrase curves are computed for the yes/no and continuation rise cases.

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:

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

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.*

2.4.2 Alignment parameter results

Figure 2 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

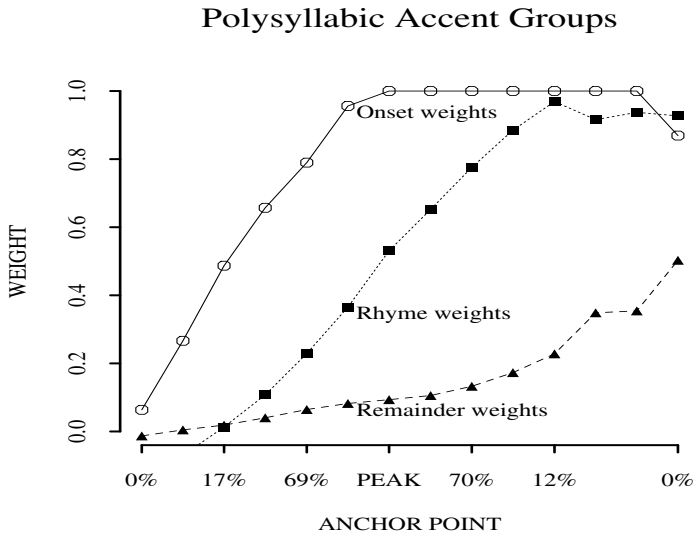

Figure 2: **Regression weights as a function of anchor point, for each of the three subintervals of the accent group. H*LL% accent type. Solid curve: onset; dotted curve: rhyme, dashed curve: remainder.**

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.

2.4.3 Additional Results

Is the time course of yes/no curves at all related to the temporal structure of the phrase-final accent group? After all, traditionally the yes/no feature is attached to phrases, not to syllables. For example, in Möbius's application of the Fujisaki model, phrase commands are not tied to accent groups (Möbius *et al.* 1993). Our data for monosyllabic accent groups strongly suggest that there is a relationship (see Figure 1). We estimated yes/no accent curves by subtracting the descending line that passes through the approximately linear initial part of the F_0 curve in the accent group up to the point where a significant rise starts; the latter point was determined by a threshold on the first derivative. As with the standard H*LL% contours, we divided the result of the subtraction by the maximum to obtain a curve ranging from 0.0 to 1.0. Thus, at the time point where the estimated accent curve starts to rise (which we call the *rise start*), the F_0 curve reaches a local minimum due to the locally descending phrase curve.

We found that the rise start could be predicted quite accurately ($r = 0.96$, $rms = 17\,ms$) from the onset duration and the sonorant rhyme duration, with respective alignment coefficients of 0.89 and 0.75. This means that the rise start is not located at some fixed or random amount of time before the phrase boundary, but varies systematically with the durations of the onset and the rhyme. In fact, the time interval between the rise start and the end of the phrase is given by:

$$t_{end\ of\ phrase} - T_{rise\ start} = (1 - 0.89) \times D_{onset} + (1 - 0.75) \times D_{rhyme} \qquad (7)$$

Thus, as measured from the end of the phrase, the rise start occurs earlier as the rhyme and the onset become longer.

Informal perceptual observations involving synthesis of polysyllabic accent groups suggest that it is important for the rise start to occur in the accented syllable, not in the phrase-final syllable. This further confirms our conclusion that yes/no rises are aligned with the phrase-final accent group in ways very similar to the alignment of declarative nuclear pitch accents.

Of course, tehse data do not exlclude an alternative account, according to which the phrase curve does not locally descend but stays flat instead, while the accent curve has an initial descending portion. In this case, it could be the local minimum and not the start of the rise whose timing is tied to the accented syllable.

Accent curves for continuation rises where estimated by subtracting a line that passes through the F_0 values at the start of the onset and at the minimum value attained before the final rise (see Figure 1).

For continuation contours, an unexpected phenomenon was found – a zero (in fact, slightly negative) correlation between the frequencies at the H* peak and the H% phrase boundary. A closer look revealed that the anchor points can be dichotomized into a pre-minimum and a post-minimum group; all correlations between groups are close to zero, all correlations within groups are strongly positive. Correlations at similarly spaced anchor points for the H*LL% and the yes/no contours were all positive, typically strongly so. One interpretation of this pattern is that continuation contours involve two component gestures – one responsible for H* and the other for H%.

2.5 Theoretical aspects of alignment model

2.5.1 Alignment model and time warping

As is often the case, a given mathematical formulation can be interpreted in ways that differ conceptually. Above, 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. Figure 3 shows how we can re-conceptualize the model in terms of *time warping of a common template.*

Panel (a) shows the percentages used to define the anchor points. Horizontal axis is the anchor point index, i in Equation 6. Panel (b) shows the alignment parameters, copied in simplified form from Figure 2. Panels (c) and (d) show the predicted anchor point locations using these alignment parameters, assuming onset and rhyme durations of 150 and 300 ms for the word *spot* and 100 and 350 ms for the word *noon*. These panels not only show the locations of the anchor points, but also display the *shape of the predicted normalized accent curves* by graphing corresponding percentage values from Panel (a) as heights of the vertical bars. These curves are called "normalized" because they range between 0.0 and 1.0.

Clearly, the two predicted normalized accent curves are similar in that they are both bell-shaped, yet one cannot be obtained from the other by uniform compression because, while the peak in "noon" is earlier, the overall horizontal extents of the accent curves are the same. It follows that there must be a *non-linear temporal relationship*, or *time warp*,

between the two curves.

Panel (e) shows the time warp that relates the anchor point locations of the two curves. Points on this curve represent the time points at which the two words reach a given anchor point. For example, the point (0.293, 0.188) represents the 11-th anchor point (which comes right after the peak).

This time warp is meaningful because there is a unique pairing (by height) of the vertical bars in Panels (c) with those in Panel (d). Now, when all curves in some set of curves can all be warped pairwise onto each other, then there must exist some common *template* onto which each curve can be warped. This is the case because the warp relationship is *transitive*: if x warps onto y and y onto z, then x warps onto z. In fact, one can take any individual curve in such a set as the template.

Finally, Panel (f) shows the time warps between the anchor point locations of the two words and anchor point indices (1-17), where we have taken the pattern in Panel (a) as the template. Note that these time warps are determined by the alignment parameters *and* the durations of the onset and rhyme.

Formally, the predicted normalized accent curve is given by:

$$\hat{F}_0(t) \;=\; P[i(t)]. \tag{8}$$

Here $i = i(t)$ 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 predicted 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 predicted 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).

2.5.2 Phonological equivalence, alignment parameters, and templates

As (Ladd 1996) states, a complete phonological description must specify how the categorical phonological elements map onto continuous acoustic parameters. In the IPO approach ('t Hart *et al.* 1990), the pitch movement inventory includes a parameterization of prototypical realizations, in particular the alignment of the movement with the segmental makeup of the syllable. This phonetic description is based on experiments that attempt

to establish the limits of variation in the realm of perception; i.e., how different can two pitch movements be acoustically and still be perceived as the same? In analogy, we ask the question: how different can two pitch accent curves be and still count as representatives of the same pitch accent type, or be derived from the same template?

We propose that what pitch accent curves in the same class have in common is that they are generated from a common template using the same family of time warp functions. They "sound the same" not only because they have the same shape (e.g., bell-shaped), but also because *they are aligned in the same way with the syllables they are associated with.* In other words, we associate a pitch accent class with an *ordered pair*:

<template><APM>

We claim that *two accent curves are phonologically distinct if, given the durations of the subintervals of their respective accent groups, they cannot be generated from the same template using the same APM.*

Consequently, the relatively small temporal shifts causing phonological changes observed by Kohler (Kohler 1990) and d' Imperio and House (d' Imperio and House 1997) could be explained by our model as follows. In both studies the segmental materials and their durations were largely left unaltered. Because the predicted accent curve that results from applying a given APM to a given template is completely determined by the segmental materials and their durations, it is impossible that the alignments of both the original and its shifted variant fit the same $< template >< APM >$ pair. Hence they must be phonologically distinct.[1]

What is somewhat difficult to grasp intuitively is that the time warp function is not constant for a given pitch accent class, but depends on the accent group sub-interval durations; what the time warp functions share is that they are generated from the same underlying APM. Our model thus provides a somewhat abstract and indirect definition for what it means for two curves to be aligned in the same way. This complexity is necessitated, of course, by the fact that all simpler alignment rules – which only applied to peak location anyway – were shown to fail.

3 Proposed pitch model

To use the predicted normalized accent curve for F_0 synthesis, three operations have to be performed.

1. The range of values has to be scaled to some appropriate range in Hz. Presumably, this scaling would reflect, among other things, the prominence of the associated pitch accent.

[1] Of course, by the same token the alignments that were all perceived as being clearly declarative (or clearly interrogative) would also require different APMs. Thus, we must add the assumption that a given accent type is associated with a probability distribution over some APM space, and that different samples drawn from the same distribution are perceived as perceptually more similar than samples drawn from distinct distributions (e.g., the interrogative and the declarative distributions). This is not any different from other multivariate situations in which categories (e.g., *big* in size judgment, *blue* in color vision, *autistic-spectrum* in clinical judgment, */ba/* in speech perception) correspond to somewhat hazy regions in multivariate spaces.

2. Once the range is scaled appropriately, the resulting curve has to be brought in line with other such curves, which requires rules for vertical placement. I.e., we must determine the location on the frequency axis of a representation of the F_0 curve which has time as horizontal axis. Vertical placement would be based on phrase-level phenomena such as declination and sentence mode.

3. Rules have to be applied for filling in gaps between successive accent curves and for resolving overlap between such curves.

Up to this point, we have refrained from using superpositional assumptions; the phrase curve was used merely as a device for computing anchor points, and we remarked that other methods that do not rely on phrase curves were rejected not on theoretical grounds but merely on the basis of our experience that they did not behave well statistically. However, we have found it difficult to do scaling, vertical placement, and successive-curve connections in any way other than in the superpositional framework. In the superpositional tradition, vertical placement is accomplished by adding accent curves to a phrase curve. Combination of successive accent curves follows as a side effect of this addition. Scaling would be accomplished by multiplying the predicted normalized accent curve with a scalar quantity.

We now discuss the details of our superpositional implementation.

3.1 Additive decomposition

In the best-known superpositional model, the Fujisaki model (Fujisaki 1983; Möbius *et al.* 1993), the observed F_0 curve is obtained by adding in the logarithmic domain three curve types with different temporal scopes: phrase curves, accent curves, and a horizontal line representing the speaker's lowest pitch level. We likewise propose to add curves with different temporal scopes, but remove the base pitch line and include segmental perturbation curves instead.

3.2 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 (Ladd 1996), 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 variability. For example, in our current work on Japanese (van Santen *et al.* 1998), phrase curves start with a rise culminating in a peak around the end of the second mora, a gentle decline until the start of the accented mora, followed by a steeper descent, and possibly terminated by a flatter region if several morae follow the accented mora.

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

3.3 Perturbation curves

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 (van Santen and Hirschberg 1994). The ratio curves (or, equivalently, difference curves in the logarithmic domain) resulting from these contrasts can be described by a rapid decay from values of about 1.30 to 1.0 in 100 ms. In our model, these curves are added in the logarithmic domain to the other curves.

3.4 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. Formally, the accent curve height parameter $H(a)$ for accent group a is given by:

$$H(a) \;=\; A[location(a)] \times B[prominence(a)] \times C[nucleus(a)] \times \cdots, \tag{9}$$

where A, B, and C are mappings that assign numbers (multipliers) to discrete levels of the arguments (e.g.: *location = initial, medial, final*). The multiplicative model is often used in segmental duration modeling. It makes the important – and not necessarily accurate – assumption of directional invariance (van Santen 1997): 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.

4 General assumptions of the model

Both our model and the model proposed by Fujisaki can be seen as special cases of a much broader superpositional or "overlay" (Ladd 1996) concept. Because discussions about the superpositional approach are often marred by focusing too narrowly on specific instances of this approach (e.g., (Ladd 1996), pp. 26–30), we feel it is important to spell out the broader assumptions of our model. This is what we hope to do in this section.

4.1 Decomposition into curves with different time courses

The key difference between our model and the Fujisaki model is that accent curves are generated by time-warping of templates vs. by low-pass filtering rectangular accent commands, respectively. Nevertheless, the two models are both special cases of the *generalized additive decomposition* concept, which states that the F_0 curve is made up by "generalized addition" of various classes of component curves:

$$F_0(t) = \bigoplus_{c \in C} \bigoplus_{k \in c} f_{c,k}(t). \tag{10}$$

C is the set of curve classes (e.g., {*perturbation, phrase, accent*}), c is a particular curve class (e.g., *accent*), and k is an individual curve (e.g., accent curve). The operator $\oplus$ satisfies some of the usual properties of addition, such as *monotonicity* (if $a \geq b$ then $a \oplus x \geq b \oplus x$) and commutativity ($a \oplus b = b \oplus a$). Obviously, both addition and multiplication have these properties.

A key assumption is that each class of curves, c, corresponds to a *phonological entity with a distinct time course*. For example, the *phrase* class has a longer scope than the *accent* class, which in turn has a longer scope than the obstruent-nonobstruent transitions with which perturbation curves are associated.

A central issue to be resolved for models in this class is which parameters of which curve classes depend on which factors. For example, in our model the alignment parameters do not depend on any phrase-level factors, and the perturbation curves are completely independent of accent status and location of the syllable containing the obstruent-nonobstruent transition.

As pointed out by Ladd (Ladd 1996), computation of these curves from observed F_0 contours is not straightforward, and is often left unspecified (e.g., (Thorsen 1983; Gårding 1983)). Fujisaki and his colleagues have been successful in estimating these curves, because of the strong assumptions of this model. We were able to fit our model because of the extreme simplicity of the recorded F_0 curves, but significant statistical problems have to be solved to apply our model to arbitrary F_0 curves. However, we have little doubt that these obstacles can be overcome. But more relevant is the point that one should not confuse these estimation difficulties with the validity of the superposition concept.

Another point raised by Ladd is that, at times, in order to obtain a good fit of the Fujisaki model phrase or accent commands have to be put in implausible locations. Of course, this point is irrelevant for the broader superpositional concept, because this result might be due entirely to some of the specific assumptions of this model, such as the exact shape of the smoothing filters.

Many issues remain to be resolved. The least-researched issue of our model is the shape of the phrase curve. While the current shape produces decent synthetic F_0 contours, we are becoming increasingly more aware of challenges, such as the necessity of multiple levels of phrasing.

4.2 Sub-interval duration directional invariance.

In the same way as addition of curves in the log domain is only a special case of a much more general decomposition principle (Eq. 10), the linear alignment model is a special case of a more general principle: the *sub-interval duration directional invariance* principle. According to this principle, for any two accent groups a and b that have the same segmental structure:

$$If\ D_j(a) \geq D_j(b)\ for\ all\ j\ then\ T_i(a) \geq T_i(b). \tag{11}$$

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):

$$T_i(a) \geq T_i(b).$$

The principle simply states that *stretching any "part" of an accent group has the effect of moving an anchor point 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/.)

An issue that needs to be addressed is the measurement of the sub-intervals. We found for the H*LL% curves that slightly different APM's were obtained depending on whether the coda was voiceless, voiced-obstruent, sonorant, or polysyllabic. It would be more elegant if the same APM's were used. There are two ways of doing this. One is to alter the definitions of the sub-interval durations, in particular the definition of where an utterance-final sonorant ends; the latter is certainly reasonable because utterance-final sonorants have no well-defined endings; we used a somewhat arbitrary energy criterion. The other is to introduce a non-linearity that would reduce the effects of very long post-accentual regions in polysyllabic accent groups. Very crudely, one could set all durations in excess of 500 ms equal to 500 ms. Any of these changes would preserve sub-interval duration directional invariance.

5 Conclusions

This paper presented data on alignment that must be accounted for by any intonation model claiming to describe both the fine and coarse details of F_0 curves. We proposed a model that accurately predicts alignment of accent curves, defined as residual curves obtained from observed F_0 contours by subtraction of a locally linear phrase curve. The model provides a very good fit. We also showed that these data cannot be accounted for

by some simple rules typical of current text-to-speech systems, which further justifies the more complicated rules embodied by the model. We described how this accent curve alignment model can be embedded in a complete superpositional model, that also incorporates phrase curves and segmental perturbation curves. Finally, we discussed generalizations of this superpositional model, and how it relates to the best-known superpositional model: the Fujisaki model.

Our model highlights the phonological importance of timing in intonation. This point has been made by many, in particular by Kohler (Kohler 1990), but has been largely ignored under the assumption that the "*" notation – used by ToBI and its predecessors to indicate with which syllable a given tone as associated – is accurate enough.

Another aspect of our model relevant for phonology has to do with the problem current intonational phonology has with mapping from the phonological level to speech. Observed F_0 curves are complicated due to intrinsic pitch effects, perturbations of post-obstruent vowels, nasality effects, presence of voiceless regions, and temporal effects of segmental durations and other factors. Together, these effects can conspire to produce spurious local maxima and minima, perturb what otherwise might have been a straight line, or create an artificial straight line. This makes it difficult to determine the locations of true peaks and lows as is required for ToBI, or the locations of short linear rises or falls as is required for the IPO approach ('t Hart *et al.* 1990). What either approach could use is a quantitative model that makes it possible to remove these effects from the observed F_0 curve; we believe that our model could play this role. What is not clear, of course, is what the relationship is between our <APM, template> pairs and the phonological entities in these approaches. So what needs further investigation is the phonological status of these <APM, template> pairs.

Finally, we discuss the relation of our work with an earlier paper by Silverman and Pierrehumbert (Silverman and Pierrehumbert 1990), in which they measured peak location (measured from stressed vowel start) in pre-nuclear high pitch accents. As in our studies, they reported effects on peak location (either in ms, or measured as a proportion of total rhyme length) of rhyme length. They also found effects of pitch accent location (nuclear vs. pre-nuclear); we have not analyzed pre-nuclear pitch accents. Our results imply that this way of measuring or normalizing peak location is problematic, because peak location is also affected by other parts of the accent group [onset duration, unstressed remainder location). We also have stated that we are not convinced that the peak should be the exclusive focus of alignment. Nevertheless, their data show that different alignment parameters are likely to be needed as a function of pitch accent location. In fact, our text-to-speech system incorporates this effect. Overall, however, our model is more in line with Bruce (Bruce 1990) who posits a more complex relationship between the F_0 contour and phonological categories than suggested by the work by Silverman and Pierrehumbert.

Acknowledgments

This paper has benefited from extensive discussions with Joseph Olive, Chilin Shih, Richard Sproat, and Jennifer Venditti. We thank Julia Hirschberg for help in the initial phases of this project. Finally, we wish to thank the reviewers – Gösta Bruce and

Cinzia Avesani – for their challenging comments.

REFERENCES

Arvaniti, A., D. Ladd, and I. Mennen. 1998. Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, Vol. 26, pp. 3–25.

Bruce, G. 1990. Alignment and compositin of tonal accents. In *Papers in laboratory phonology I: Between the grammar and physics of speech* (Kingston, J. and M. E. Beckman, editors), pp. 107–114, Cambridge, UK: Cambridge University Press.

Caspers, J. 1994. *Pitch movements under time pressure*. PhD thesis, Leiden University.

d' Imperio, M. and D. House. September 1997. Perception of questions and statements in Neapolitan Italian. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*, (Rhodes).

Fujisaki, H. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech* (MacNeilage, P. F., editor), pp. 39–55, New York: Springer.

Gårding, E. 1983. A generative model of intonation. In *Prosody: Models and measurements* (Cutler, A. and D. R. Ladd, editors), pp. 11–25, Berlin: Springer.

Kohler, K. 1990. Macro and micro F0 in the synthesis of intonation. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (Kingston, J. and M. Beckman, editors), pp. 115–138, Cambridge: Cambridge University Press.

Ladd, D. 1996. *Intonational phonology*. Cambridge University Press, Cambridge, UK.

Möbius, B., M. Pätzold, and W. Hess. 1993. Analysis and synthesis of German F0 contours by means of Fujisaki's model. *Speech Communication*, Vol. 13.

Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Silverman, K. and J. Pierrehumbert. 1990. The timing of prenuclear high accents in English. In *Papers in laboratory phonology I: Between the grammar and physics of speech* (Kingston, J. and M. E. Beckman, editors), pp. 72–106, Cambridge, UK: Cambridge University Press.

't Hart, J., R. Collier, and A. Cohen. 1990. *A Perceptual Study of Intonation*. Cambridge University Press, Cambridge UK.

Thorsen, N. 1983. Two issues in prosody of standard Danish. In *Prosody: Models and Measurements* (Cutler, A. and D. R. Ladd, editors), pp. 27–38, Springer-Verlag.

van Santen, J. and J. Hirschberg. 1994. Segmental effects on timing and height of pitch contours. In *Proceedings ICSLP '94*, pp. 719–722.

van Santen, J., B. Möbius, J. Venditti, and C. Shih. 1998. Description of the Bell Labs Intonation System. In *Third ESCAWorkshop on speech synthesis*, (Jenolan Caves, Australia).

van Santen, J. September 1997. Prosodic modeling in text-to-speech synthesis. In *Proceedings of Eurospeech-97*, (Rhodes).

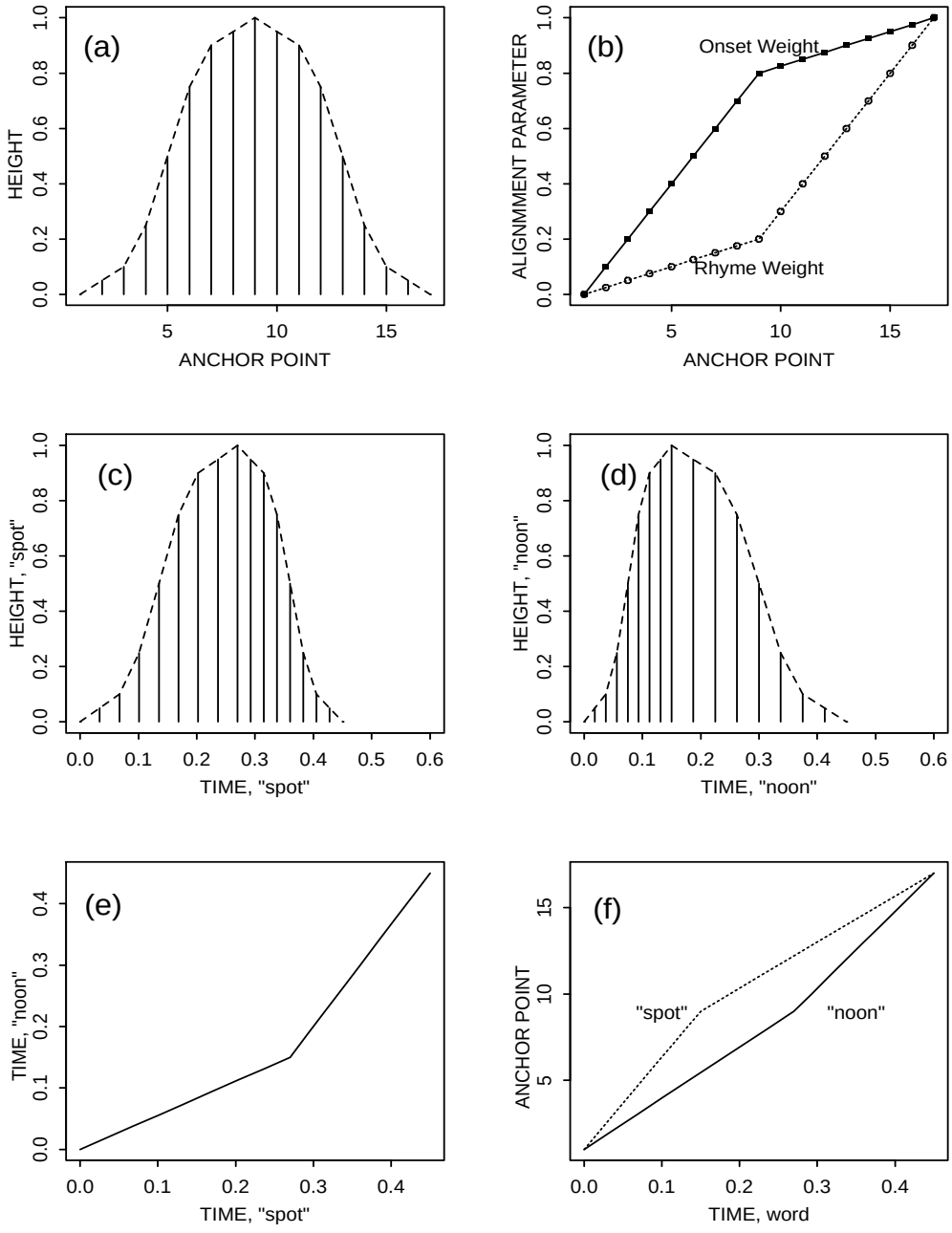

Figure 3: **Relationship between alignment parameters and time warping. (a): Template, defined as an array of fractions. (b): Hypothetical alignment parameters. (c): Predicted anchor points for the word "spot", using these alignment parameters and an onset duration of 300 ms and a rhyme duration of 150 ms. (d): Predicted anchor points for the word "noon", (onset duration: 100 ms, rhyme duration: 350 ms.) (e): Time warp of "noon" onto "spot", showing locations of corresponding anchor points. (f): Time warps of "noon" and "spot" onto the template.**