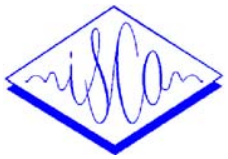

Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

Quantitative Modeling of Pitch Accent Alignment

Jan P. H. van Santen

Center for Spoken Language Understanding
OGI School of Science & Engineering at the Oregon Health & Science University
vansanten@ece.ogi.edu

10.21437/SpeechProsody.2002-15

Abstract

This paper poses two interrelated questions. (i) What aspects of pitch movement and the segmental stream are aligned? (ii) What does it mean for these aspects to be aligned? Our basic claim is that these aspects are unlikely to be found at the acoustic surface level, and that alignment itself involves an abstract relationship rather than, for example, the simple coincidence of certain discrete pitch events (e.g. peaks) with discrete segmental events (e.g., syllable boundaries.) This abstract relationship consists of a mathematical mapping between intonational and segmental trajectories that is invariant within a given phonological category and "speech state." We present a conceptual framework that makes this claim more precise, and illustrate the framework by discussing in detail a specific model, the *Linear Alignment Model* (A Model of Fundamental Frequency Contour Alignment", in A. Botinis, Ed., *Intonation: Analysis, Modeling and Technology*.)

1. Introduction

To use a deliberately vague formulation, alignment refers to the temporal relationship between pitch events and segmental events. As pointed out by Xu [13], this relationship is often assumed to be loose: the defining characteristic of pitch movement (e.g., a local peak) occurs during a possibly briefly after a syllable. However, Xu [13] presents compelling evidence that alignment is more constrained. Constraints mentioned are of a neurophysiological nature: Built-in limits on F_0 velocity, and the nervous system's tendency towards synchronization when complex tasks are performed involving multiple motor programs.

Evidence for an even tighter coordination between pitch events and segmental events comes from perceptual results that indicate that small changes in alignment can cause changes in meaning [5, 3, 2]. These perceptual findings suggest that speakers are able to control alignment at this level of precision. It seems unlikely that such constrained behavior is purely due to the neurophysiological constraints discussed by Xu. Rather, these constraints may be dictated by perceptual concerns [2].

These perceptual results are challenging against the background of another set of results showing large variability due to segmental and durational influences [8]. In their studies, a speaker produced more than 2,000 utterances grouped by pitch accent type, varying a single target word within each type. All other factors were constant, including speaking rate and affective state. As measured in terms of peak location, alignment varied over a larger range than the range studied by [5, 3, 2]. Yet, these variations did not cause variation in perceived meaning or phonological class.

These two sets of results present a paradox that needs to be addressed by any conceptualization of alignment: On the one hand, keeping the segmentals roughly the same, small changes in alignment can cause a change in perceived meaning, whereas large changes in alignment due to segmental and durational factors do not necessarily cause a change in perceived meaning.

The paradox raises two questions:

- What pitch events and segmental events are involved in alignment? Are they indeed such things as peaks and low points? Or do we need to search for deeper components that underly surface features?
- Independently of whether surface or "deep" features are involved in alignment, what is it that stays invariant across pitch contours associated with a given phonological category?

Apparently, there must exist some deeper invariance in the space of pitch contours associated with a given phonological category across segment sequences. This paper claims that the best way to capture this invariance is through a quantitative intonation model. More specifically, in order to abstract away from a surface F_0 contour that is unavoidably polluted by segmental influences and voicing irregularities, we argue that one should consider models that are able to dissociate these influences and irregularities from the primary features of interest – accentuation and phrasing.

The argument in favor of quantitative modeling has two sources. One is based on the claimed validity of a particular quantitative model – the *Linear Alignment Model* [9]. This is the bulk of the argument in this paper. However, there is a separate argument based on the following speculation about the abstract nature of speech production. We speculate that the ultrafast and extremely complex operation of muscles involved in speech, including those for generation of pitch (i.e., larynx) and for generation of segmentals (which, besides everything else, of course includes the larynx), involves a *coordinated state-dependent system* in which, within a given "speech state", the coordination between these muscles is quite tight – much tighter than what is dictated by the neurophysiological constraints discussed by Xu [13]. It may very well be the case that this tight coordination is a neural necessity, given the speed and complexity of the motor task. It may additionally be the case that each coordinative pattern (i.e., the overall trajectory space for a given speech state) is optimized to achieve certain perceptual goals.

In the remainder of the paper, we first provide a detail description of the Linear Alignment Model, and then discuss its implications for alignment. Finally, we generalize this account in terms of a broader class of models than the Linear Alignment Model.

2. Linear Alignment Model

The Linear Alignment Model is a superpositional model that pays particular attention to alignment. We first describe the general superposition concept, and then provide details of the model

2.1. General Concept of Superposition

We first briefly describe the seminal example of superpositional models, the *Fujisaki model* [4]. In the Fujisaki model, the intonation contour for a given phrase is obtained by addition (in the logarithmic domain) of a *phrase curve* and zero or more *accent curves*. The phrase curve has the temporal scope of a phrase, and is completely specified by a start and end time, and by sentence mode (declarative, interrogative, etc.) In other words, the phrase curve is unaffected by whether any syllables are accented or where in the phrase pitch accents occur. The phrase curve is generated by applying a second-order linear filter to impulses called *phrase commands*.

Accent curves (at least in versions of the model that have been applied to Japanese and English) have an up-down pattern, starting and ending at a value of zero. They correspond to accented syllables, and have a temporal extent that roughly coincides with a syllable or sequence of syllables; "roughly", because although the starting point of an accent curve coincides with the start of an accented syllable, the end point does not necessarily correspond to any syllable boundary. The parameters of an accent curve (start time, end time, height) are independent of phrasing. Accent curves are generated by applying a filter to rectangular functions called *accent commands*.

This example illustrates the key aspects of the general superpositional approach, which we now discuss using the same formalism as in [9].

In the general superpositional approach, the intonation curve is viewed as the *generalized addition* (addition in the log domain is an example of generalized addition) of underlying *component curves* that belong to one of several *component curve classes*. These classes differ in their temporal scope and in the type of linguistic entity they are tied to. Formally,

$$F_0(t) \;=\; \bigoplus_{c \in C} \bigoplus_{k \in c} f_{c,k}(t). \qquad (1)$$

Here, $F_0(t)$ is pitch in Hz at time t, C is the set of curve classes (e.g., {*phrase, accent*}), c is a particular curve class (e.g., *accent*), and k is an individual curve (e.g., a specific accent curve). The operator $\bigoplus$ 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 central assumption is that each class of curves, c, corresponds to a *phonological entity class with a distinct temporal scope*. For example, the *phrase* class has a longer scope than the *accent* class. This assumption provides for a *conceptually clear link between linguistic (and para-linguistic) control factors and surface features of the pitch contour*. Whether the link as proposed by the class of superpositional models is in fact accurate is a separate matter; the key point is that superpositional models illustrate how such a link can be conceptualized.

2.2. Empirical Findings Underlying the Linear Alignment Model

We first describe some of the empirical findings that formed the basis for the Linear Alignment Model, and then describe the model proper. We briefly summarize here the main findings of [8] (see also: [10, 9].) Their experiments involved a single speaker, producing target words differing in segmental and syllabic contents, in systematically varied sentence frames. Unless specified otherwise, in most experiments discussed here the target word always carried the nuclear pitch accent: single intonational phrases, with a single H* pitch accent, a low phrase accent, and low boundary tone. Initial measurements focused on peak location, which was measured in several ways, including from the start of the accented syllable, from the start of the first sonorant of the accented syllable, and from the start of the nucleus. Peak location was also measured either in ms or in relative terms, as a fraction of the accented syllable, of the accented syllable rhyme (or, rather, *s-rhyme*, defined as the interval beginning with the start of the last sonorant in the onset (or vowel start if the onset has no sonorants) and that ends at the end of the last sonorant in the syllable), and of the combined duration of the accented and unaccented syllables.

Among the key findings were these.

(i) Peak placement (in ms and relative) depended on the phonetic classes of the onset (C_o) and coda (C_c).

(ii) Peak placement (in ms) also depended on the durations of the segments.

(iii) The joint effects of phonetic class and durations could be predicted quite well with the model (for phrase-final accented syllables):

$$\begin{aligned} T_{peak}(D_{onset},\, D_{s-rhyme};\, C_o,\, C_c) \;=\; & \\ \alpha_{C_o,C_c} \times D_{onset} & \\ +\; \beta_{C_o,C_c} \times D_{s-rhyme} \;+\; \mu_{C_o,C_c} & \qquad (2) \end{aligned}$$

According to this model, peak time for a syllable whose onset duration is D_{onset} and s-rhyme duration is $D_{s-rhyme}$ is a weighted combination of these two durations plus a constant, which, like the weights, may depend on C_o and C_c.

(iv) When the accented syllable was followed by one or more deaccented syllables, the model was amended as follows:

$$\begin{aligned} T_{peak}(D_{onset},\, D_{rhyme},\, D_{rest};\, C_o) \;=\; & \\ \alpha_{C_o} \times D_{onset} \;+\; \beta_{C_o} \times D_{rhyme} & \\ +\; \gamma_{C_o} \times D_{rest} \;+\; \mu_{C_o,C_c} & \qquad (3) \end{aligned}$$

Here, D_{rest} is the combined duration of the deaccented syllables.

(v) The values of the parameters α, β, γ, and μ were instructive. We found that:

$$1 \;>\; \alpha \;>\; \beta \;>\; \gamma \;>\; 0 \qquad (4)$$

and

$$\mu \;=\; 0 \qquad (5)$$

This contradicts various simple models of peak placement, such as: *fixed percentage into syllable* (because $\alpha \neq \beta$), *fixed percentage into the vowel or s-rhyme* (because $1 \;>\; \alpha$), and *fixed*

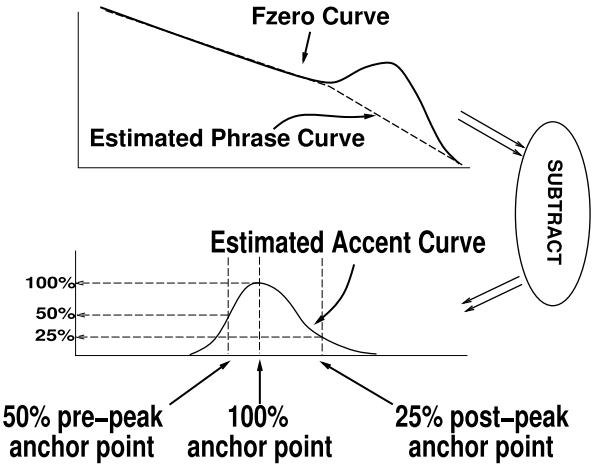

Figure 1: *Estimation of accent curves and anchor points. The top panel shows the generation of the phrase curve; the bottom panel shows the shape of the estimated accent curves and the computation of anchor points.*

ms amount into the syllable (because $\alpha > \beta > \gamma > 0$). The finding that $\gamma > 0$ is of interest, because it shows that peak placement in accented syllables followed by at least one unaccented syllable is influenced by the durations of the latter. This suggests that not the accented syllable but *the (left-headed, or trochaic) foot is the unit with which these pitch accents are associated.*

In addition to these findings on peak placement, also effects were found of intrinsic pitch and of segmental perturbation (defined as effects during the initial 50-100 ms of a vowel preceded by an obstruent). To summarize:

(i) The effects of segmental perturbation could be described as a fast decaying effect, with an initial value of at least 20 Hz or 15%.

(ii) The size of these effects was completely independent of the accent status of the syllable and of the location in the phrase.

(iii) Intrinsic pitch effects were found, but only in accented syllables.

All results reported sofar were based on peak location, for reasons of convenience and tradition. However, we wanted to understand alignment of the entire trajectory. Towards this end, we used a procedure in which the model for peak location (i.e., Eqs. 2 and 3) was extended using a *superpositional approach*. Because of the extreme simplicity of the obtained pitch contours, it was easy to draw a line from the start of the pitch accent to the end of the phrase. This line, which could be considered as a local estimate of the local phrase curve, was then subtracted from the observed pitch curve (Fig 1.)

Subtraction produces a curve that rises from zero to a peak value, and then returns to zero. This process allows us to then characterize the shape of the rise-fall pattern using the concept of "anchor point". For each relative height value or *anchor value* we can find the corresponding point on the time scale. This point is called the *50% pre-peak anchor point.* Note that the peak location itself is simply the *100% anchor point*.

Given a set of anchor values, the spacing pattern of the corresponding anchor points *completely characterizes the shape of the rise-fall pattern.* This way of representing curve shape differs from other methods, such as fitting a polynomial function or characterizing the curve in terms of fall-start, point-of-steepest ascent, and the like.

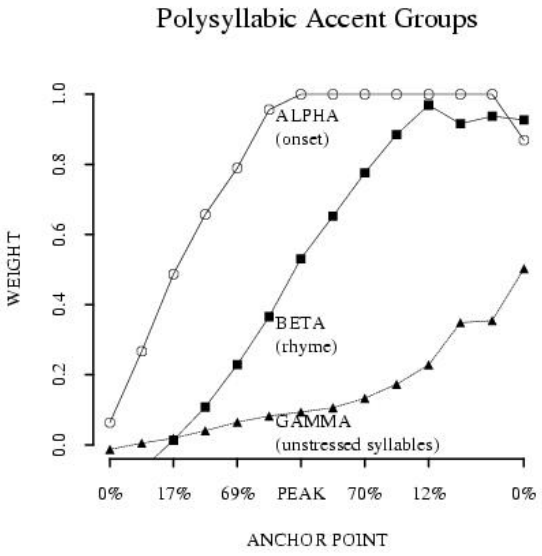

Figure 2: *Alignment parameters.*

The generalization of Equations 2 and 3 to accomodate the anchor point concept is obvious (A refers to the A-th anchor point):

$$T_A(D_{onset},\ D_{s-rhyme};\ C_o,\ C_c) = \alpha_{C_o,C_c,A} \times D_{onset} + \beta_{C_o,C_c,A} \times D_{s-rhyme} + \mu_{C_o,C_c,A}. \quad (6)$$

and

$$T_A(D_{onset},\ D_{rhyme},\ D_{rest};\ C_o) = \alpha_{C_o,A} \times D_{onset} + \beta_{C_o,A} \times D_{rhyme} + \gamma_{C_o,A} \times D_{rest} + \mu_{C_o,C_c,A}. \quad (7)$$

We call the ensemble of parameters α, β, γ, and μ *alignment parameters*. Figure 2 shows their values for polysyllabic contexts (i.e., the accented syllable was followed by at least one unaccented syllable). The parameter μ was dropped, because it was uniformly found to be 0.

In what follows, we describe how Eqs. 6 and 7 can be used for the generation of accent curves within a superpositional framework.

2.3. Curve Classes

The Linear Alignment Model uses three curve classes: Phrase Curves, Segmental Influence Curves, and Accent Curves. In some versions, such as for Japanese [12], further curve classes are added such as UA curves (associated with sequences of zero or more lexically unaccented words followed by a lexically accented word or higher-level phrase boundary. Figure 3 shows that this analysis has the interesting feature of modeling the events surrounding the accented mora as a rise-fall accent curve pattern (on the bottom of the figure) superposed on

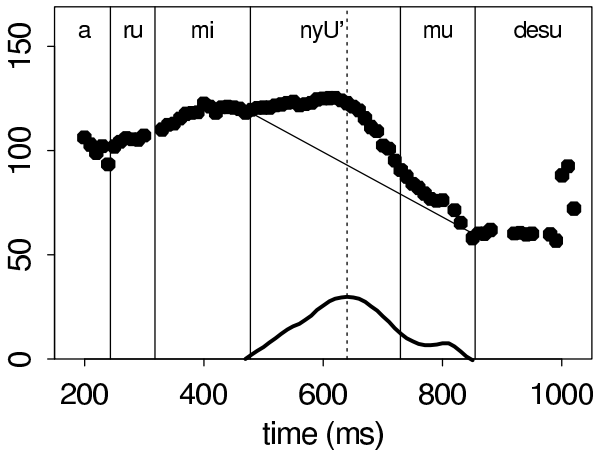

Figure 3: *Analysis of a UA-group F_0 contour. The bottom curve is the accent curve that accounts for the late fall in the accented mora when a local UA curve (thin line) is subtracted from the F_0 curve. From: Venditti & van Santen, "Japanese Intonation Synthesis using Superposition and Linear alignment", Proc. ICSLP 2000*

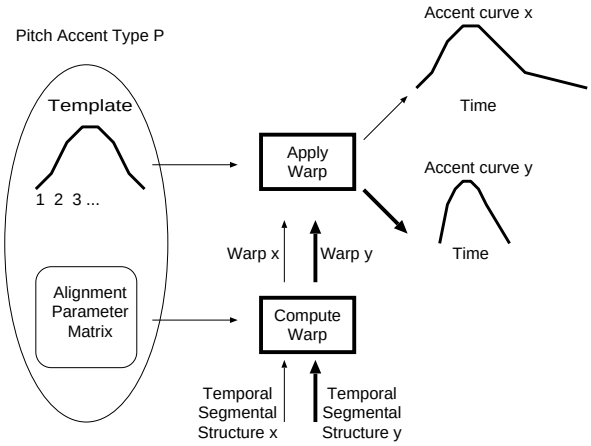

Figure 4: *Flow diagram of accent curve generation. Temporal pattern* **x, y** *are combined with a pitch-accent-specific alignment parameter matrix to form time warps,* warp x, warp y, *that are applied to a pitch-accent-specific template to generate accent curves.*

a locally sharply declining UA curve; the net effect is a fall in the F_0 curve towards the end of the accented mora. However, the key phonological event is not the fall itself – which is just a by-product of the relative time course of these two underlying curves – but the *underlying* rise-fall pattern on the accented mora itself.

2.3.1. Phrase Curves

Whereas the Fujisaki model makes strong assumptions about the phrase curve, in Linear Alignment Model the shape of the phrase curve is essentially unconstrained except for the broad assumption that it should be quite smooth over long time stretches. In actual applications, such as in the Bell Labs TTS system [10], it consists of two quasi-linear segments, one starting at the phrase onset and ending at the onset of the nuclear pitch accent and the other segment continuing this segment to the end of the phrase.

2.3.2. Segmental Influence Curves

The segmental perturbation curves reflect intrinsic pitch, effects on vowels of preceding obstruents, and lowering in sonorants. These are modeled by either additive (post-obstruental perturbation) or multiplicative (intrinsic pitch) parameters.

2.3.3. Accent Curves

In most applications of the Linear Alignment Model, accent curves are associated with trochaic, or left-headed, feet, defined as a sequence of one or more syllables in which only the first syllable is accented. A foot is terminated either by the next accented syllable or by a phrase boundary. No provisions are made for secondary stress. Of course, this was based on [8], where we focused on syllables carrying nuclear pitch accents and where these syllables where followed by varying numbers of unaccented syllables. We could have written, equivalently, that we were varying trochaic foot length.

In the Linear Alignment Model, accent curves are generated in a way that differs fundamentally from the Fujisaki model. Specifically, we make use of Eqs. 6 and 7 to generate accent curves *from templates via parameterized time warp functions.* As we shall see, this will ensure that the exact time course of these accent curves will closely approximate the natural (in the cases analyzed, rise-fall) patterns that were analyzed in [8], or, put differently, that the temporal coordination between the rise-fall pattern and the the associated sequence of segments or syllables mimics that found in natural speech.

For a pitch accent type P, we define its template as a sequence of anchor values $T_P = < P_1, \cdots, P_n >$. These anchor values describe the archetypical shape of P. For example, for a pitch accent type associated with a rise fall pattern, the template might be:

$$T_P = <0,\ 0.05,\ 0.2,\ 0.8,\ 0.9,\ 1.0,\ 0.9,\ 0.8,\ 0.2,\ 0.05,\ 0.0>$$

Also associated with P is an alignment parameter matrix M_P that contains all values of α, β, γ. Given a rendition of trochaic foot with durations of D_{onset}, $D_{s-rhyme}$ (or D_{rhyme} and D_{rest}), the A-th anchor point is located on the time axis as indicated by Eqs. 6 and 7, and its corresponding frequency value is P_A (ultimately to be multiplied by an *amplitude parameter* that reflects the degree of emphasis). Figure 4 shows the flow diagram of this operation.

A corollary of the above is the following: All accent curves for pitch accent type P share that they are generated from a common template and alignment parameter matrix. They differ from each other solely because the durations D_{onset}, etc., differ. In other words,

$$\begin{aligned} Curve\ = \\ f(&Alignment\ Parameters,\ Template, \\ &Temporal\ Segmental\ Structure) \end{aligned} \tag{8}$$

where "Temporal Segmental Structure" is defined as the sequence of the phonetic classes and durations of the segments that make up a foot. (we presume here that segmental effects are reducible to effects of segmental classes.)

3. What is Invariant about Alignment?

We now return to the basic questions raised in the introduction: *What pitch events and segmental events are involved in alignment*, and; *What stays invariant across pitch contours associated with a given phonological category?* We first provide answers provided by the Linear Alignment Model, and then generalize these answers.

3.1. The Linear Alignment Model and Alignment

According to the Linear Alignment Model, the answer to the first question is as follows. On the segmental side, the Linear Alignment Model opts for sub-intervals of trochaic feet. There is some arbitrariness in this; the only specific segmental event that the Linear Alignment Model makes strong assumptions about is the start of the accented syllable. (There is increasingly stronger evidence for the special status as an anchor point of the start of the accented syllable: [2, 1]) On the pitch side, the situation is notably different from the usual approach in terms of peaks and lows. Now, the "events" are not specific points, but are trajectories. Moreover, they are not trajectories of raw F_0 contours, but of accent curves, defined as foot-long deviations from an underlying phrase curve from which also segmental disturbances have been removed.

This account of alignment has the following advantages over accounts in terms of surface events such as local pitch maxima or minima:

(i) Local maxima can result from segmental perturbations. In a word such as "sit", the pitch values in the initial part of the vowel can exceed those at the true peak location later in the vowel.

(ii) When there is a steep underlying phrase curve, such as in Figure 3, there may hardly be a local maximum even if the underlying accent curve does have a rise-fall pattern.

(iii) In polysyllabic trochaic feet, the peak is often around the syllable boundary. Whether it is located on one side or the other side of the boundary is purely the result of the segments and their durations, and does not carry implications for meaning or intention.

(iv) Under these same circumstances, peaks can be "hidden" when the segments surrounding the boundary are not sonorants.

Concerning the second question: A given pitch accent is defined by the combination of a template and an alignment parameter matrix (Figure 4). Together, these define a mapping from temporal segmental structures on the one hand to accent curves on the other hand. In other words, they define how accent curves and temporal segmental structures are coordinated. According to the Linear Alignment Model, the change in perceived phonological category due to small displacements is due to curves that cannot have been generated by the same Template + Alignment Parameter Matrix combination (and hence pitch accent class), because the displacements were generated *while keeping the temporal segmental structure the same*. Thus, even though the shape, and hence potentially the underlying templates, were the same, the alignment parameters cannot also have been the same. In terms of Eq. 8, using c for $Curve$, ts for $Temporal\ Segmental\ Structure$, M for $Alignment\ Parameters$, and T for $Template$: $c \neq c'$ and $ts = ts'$ implies either $M \neq M'$ or $T \neq T'$. Of course, in van Santen and Hirschberg's production studies [8], always $ts \neq ts'$

3.1.1. The concept of "target"

In earlier work on non-tone languages, pitch contours were described in terms of movement between "tonal targets", defined as points in Time × Frequency space. On the phonological side, abstract tones or tone combinations were the basic building blocks, and these were seen to map in a relatively straightforward – yet rarely precisely specified – manner on the acoustics [6]. Xu [13] proposes more complex targets, in the form of pitch movements, or trajectories in Time × Frequency space. Also in certain approaches to intonation synthesis, such as the IPO approach [7], not points but trajectories in the form of line segments are used. The Linear Alignment Model does not have targets in the sense that a speaker is trying to achieve a particular pitch value or pitch value trajectory.

Given a specific temporal segmental structure, any reasonable stochastic version of the Linear Alignment Model would allow for substantial variation in local phrase curve shapes and of accent curves amplitudes. Even in the tightly controlled recordings analyzed in [8], the amplitude of the accent curves would vary considerably. (Interestingly, certain other aspects of the F_0 curves, such the final boundary town, were remarkably constant.)

Across temporal segmental structures, pitch contours are even more variable: The "target" is the *space of trajectories* defined by a template and an alignment parameter matrix. What stays constant is not pitch values, but abstract dynamic patterns.

3.2. Speech-state dependent alignment

An important shortcoming of the Linear Alignment Model is that it is too constraining. Obviously, the recordings analyzed in [8] are tightly controlled: The same speaker, strictly supervised, reads sentences with the same sentence frame in succession. It is thus no surprise that one particular model with a fixed set of parameters provided an excellent fit. At the same time, it provides evidence that speakers are capable of such highly constrained speech, even though neither meaning, context, nor neurophysiology dictated such constrained speech behavior.

As in the Introduction, we invoke here the speculative concept of *speech state.* In the Introduction, we speculated that the coordination of pitch movement and segmentals involves a state-dependent system. In terms of the Linear Alignment Model, this can be represented as a transformation of the alignment parameter matrix (Figure 5).

In the Lucent Technologies Bell Labs Multilingual TTS system [11] which uses the Linear Alignment Model for most languages, this is accomplished essentially with matrix multiplication. This enables making global changes in the synthesizer's behavior, such as the average peak placement, the degree of overlap between successive accent curves, and the like.

Abstracting from the specific way in which the Linear Alignment Model embodies the general concept of speech state dependent alignment, clearly some version of this concept is useful. Constraints on alignment have multiple sources and types. "Hard" constraints include those set by the limits of the speech production apparatus and by perceptual classifiability (e.g., in American English, a yes/no question cannot be compellingly conveyed by sharp sentence-final lowering). Within these constraints, a wide range of variation is possible. We conceive of this variation not as random, but as state dependent:

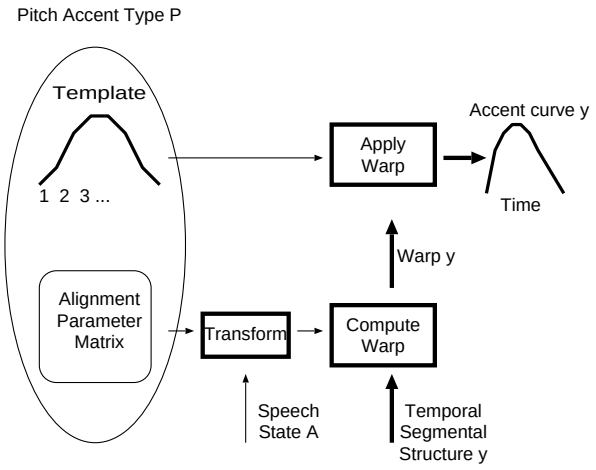

Figure 5: *Flow diagram of accent curve generation, with speech state dependent transformation of alignment parameters.*

Within a given state, there is tight alignment; different speech states may convey different affective states, speaking rates, and other para-linguistic factors. However, states may differ in major ways both within and across speakers.

4. Discussion

The main goal of this paper is to make the point that the study of alignment requires quantitative modeling. Towards that end, we introduced the Linear Alignment Model, and showed how it could shed light – be it a speculatively – on some of the more puzzling questions about alignment: *What is it that is aligned*, and *what is it that is invariant?*.

The model shows that one can discuss alignment with perfect clarity without referring to such entities as pitch peaks, rises, and targets. The model views alignment as a mapping between the temporal segmental structure of a foot and the trajectory of a local excursion. The concept of "local excursion" presupposes a superpositional framework, which may be too confining for some. The model most certainly does not view alignment as the coincidence of specific surface pitch events and specific segmental anchors, with the possible exception of the accented syllables start.

A model of this type raises interesting research questions. First, the Linear Alignment Model is incomplete to the extreme. It has only been developed to include a small number of pitch accent types. In addition, no attempt has been made to cover tone languages.

Second, it would be of interest to measure changes in alignment parameters as a function of factors such as speaking rate, affective state, and phrase-wide factors such as position in a paragraph or sentence mode. Alignment parameters – which are easy to estimate in practice, using multiple regression – will provide a more valid (because free of segmental influences) and richer (because it captures the entire trajectory, not just one point) measure than the usual peak location measures.

Third, the superposition concept is a controversial framework, yet few if any attempts have been made to test the framework itself. This is easier said than done, however, because a formulation as in Eq. 1 is quite general and may not have obvious testable predictions. A successive narrowing down may lead to such predictions, however. For example, the Linear Alignment Model makes a strong asymmetry prediction: The number of unstressed syllables preceding an accented syllable should have far less effect on pitch movement on that syllable than the number of following unstressed syllables.

5. References

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

[2] D'Imperio, M. Language-specific and universal contraints on tonal alignment: the nature of targets and "anchors". In *Proceedings of Speech Prosody 2002* (Aix-en-Provence, 2002).

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

[4] 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.

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

[6] Pierrehumbert, J. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, September 1980. Distributed by the Indiana University Linguistics Club.

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

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

[9] van Santen, J., and Möbius, B. A model of fundamental frequency contour alignment. In *Intonation: Analysis, Modelling and Technology*, A. Botinis, Ed. Cambridge University Press, 1999. In press.

[10] van Santen, J., Shih, C., and Möbius, B. Intonation. In *Multilingual Text-to-Speech Synthesis*, R. Sproat, Ed. Kluwer, Dordrecht, the Netherlands, 1997.

[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.

[12] Venditti, J., and van Santen, J. Japanese intonation synthesis using superposition and linear alignment models. In *Proceedings ICSLP* (Beijing, China, 2000).

[13] Xu, Y. Articulatory constraints and tonal alignment. In *Proceedings of Speech Prosody 2002* (Aix-en-Provence, 2002).