Toward measurement of pitch alignment

Jan P.H. van Santen, Esther Klabbers & Taniya Mishra

This paper discusses theoretical and practical issues underlying the measurement of pitch alignment. We define the alignment concept as the relationship between pitch trajectories and articulatory/acoustic trajectories. This concept is formalized within a general superpositional framework, according to which a pitch curve is viewed as the sum of component curves, such as phrase curves, accent curves, and segmental perturbation curves. According to a special case of the general superpositional concept, the Linear Alignment Model, a given intonational-phonological pitch accent class can be characterized as the combination of an underlying accent template (that represents the basic shape of the pitch excursion, e.g., rise, rise-fall) and an alignment parameter matrix (that specifies how to warp this template to be properly aligned with the segmental stream with which a pitch accent is associated, taking into account the segmental/durational structure of this stream).

Traditional measurement of alignment is customarily based on the surface-point-to-surface-point approach (P-P approach). In this approach, the time interval (in ms or as a percentage of some structural unit such as the syllable) is measured between pitch points (i.e., points on the surface pitch contour, such as local F_0 minima and maxima) and segmental points (i.e., points in the segmental stream such as boundaries between segments, syllable constituents, or syllable boundaries). A special case of the P-P approach is the search for segmental points that serve as anchors for pitch points.

We show how the Linear Alignment Model can account for the systematic dependency of pitch point timing on the segmental/durational structure of the segmental stream. Specifically, we show how apparent changes of alignment as measured by the P-P approach, resulting from some independent variable, are in fact predictable via the model as direct consequences of the effects of the independent variable on the segmental/durational structure, and thus may not be changes in alignment at all.

We also show how the model can account for phonological-perceptual changes associated with small changes in alignment in combination with unchanged segmental/ durational structure and pitch accent shape.[*]

1. Introduction

During the past decade, pitch alignment has been receiving increasingly more attention (e.g., (Ladd 1983; Arvaniti *et al.* 1998; Ladd 2000, 1999; D'Imperio 2002; Prieto *et al.* 1994; Venditti & van Santen 2000; van Santen & Hirschberg 1994)). Yet, close inspection of the literature shows that its measurement is a far from solved problem.

Rivista di Linguistica 18.1 (2006), pp. 161-187 *(ricevuto nel settembre 2006)*

We distinguish between the theoretical conception of alignment and approaches for its measurement. Conceptually, we view alignment as an abstract relationship between pitch trajectories and articulatory/acoustic trajectories. This relationship, which is presumably dictated by a host of physiological, communicative, and other constraints, is what makes a given phonological intonational event perceptually invariant across a range of contexts where it is intended to be perceived as invariant. For example, the F_0 curves for H* pitch accents in renditions of the words *sir* and *surface* will be measurably different simply as a result of the segmental and durational differences between these two words, even when they are perceived and intended to be the same. The challenge is: How to characterize what stays invariant in the relationship between the articulatory/acoustic trajectories on the one hand and the pitch trajectories on the other hand?

The traditional approach toward measuring alignment is based on the surface-point-to-surface-point (P-P) method. In this approach, the time interval (in ms or as a percentage of some structural unit such as the syllable) is measured between pitch points (i.e., points on the surface pitch contour, such as local F^0 minima and maxima) and segmental points (i.e., points in the segmental stream such as boundaries between segments, syllable constituents, or syllables). The adequacy of the P-P measurement method depends on whether one believes that the relationship between pitch trajectories and articulatory/ acoustic trajectories can be adequately characterized by pitch points and segmental points, or whether a more complex characterization is needed.

We will argue in this paper that the P-P method is indeed too simple, and that alignment measurement based on a specific superpositional pitch model, the Linear Alignment Model, is better able to characterize the relationship between pitch trajectories and articulatory/acoustic trajectories. Specifically, we show how apparent changes of alignment as measured by the P-P approach, resulting from some independent variable, are in fact predictable via the model as direct consequences of the effects of the independent variable on the segmental/durational structure, and thus may, in fact, not be changes in alignment.

The outline of the paper is as follows. Sections 2 and 3 provide a general characterization of the superpositional concept and an instantiation of this concept, the Linear Alignment Model, respectively. Section 4 analyzes the implications of the model the measurement of alignment. Finally, Section 5 presents new applications of the model.

2. General Characterization of Superpositional Models

Superpositional models, exemplied by the Fujisaki Intonation Model (Fujisaki 1983) and the Linear Alignment Model (van Santen & Möbius 1999), hold that the pitch curve can be thought of as being composed (via an addition-like operation) of component curves that belong to multiple curve classes.

We briefly describe the Fujisaki model (Fujisaki 1983). In this 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; 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 shapes called accent commands.

The Fujisaki model illustrates the key aspects of the general superpositional model, which we now discuss using the same formalism as in (van Santen & Möbius 1999).

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,

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

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 $\oplus$

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 (as in the Fujisaki model) have these properties.

A key additional assumption is that each class of curves, c, has a distinctive time scale and is the resultant of different underlying articulatory processes.

Concerning time scale, a phrase curve has a time scale ranging from a single word to many words, an accent curve is associated with a foot and has a temporal extent of a few hundred ms, and a curve that models segmental perturbation (absent in the Fujisaki model, but introduced in the Linear Alignment Model, below) has a sub-segmental time scale. We posit that accent curves and phrase curves may be controlled by different underlying muscle systems. For example, Fujisaki provides evidence for a specific link between accent curves and phrase curves and the quasi-independent actions of different parts of the cricothyroid muscle (Fujisaki 1988, 1995). We also posit that the processes responsible for segmental perturbations are, again, quasi-independent from the processes responsible for accent and phrase curves.

This assumption reflects a bias in our modeling that emphasizes a close tie of the model to acoustics and articulation, potentially at the expense of the simplicity of the linguistic interpretation of the model. In the same way as an articulator such as the tongue, or in fact the larynx itself, can serve both prosodic and segmental purposes and thus cuts across linguistic categories and levels, we should not necessarily assume that there is a one-to-one relationship between these curve classes and classes of phonological entities and their associated features. Whether such a relationship exists is an empirical issue. To illustrate this point, the Bell Labs speech synthesis implementation of the Linear Alignment Model (van Santen, Shih & Möbius 1997) unabashedly models sentence mode (e.g., yes/no question) via a rising accent curve. The reasoning behind this was based on the informal observation that the yes-no rise pattern had similar alignment to the phrase-final foot as rise-fall patterns, with the rise starting at the beginning of the accented syllable in both cases independently of the presence and number of unstressed post-accentual syllables. The local minimum, which is generally considered an important feature of yes-no question rises, is a side effect of the initial shallowness of the rise in combination with the sharp descent of the phrase curve at the start of the syllable carrying the nuclear pitch accent.

A similar message about the complexity of the relation between curve classes, phonological entities, and their features, could be extracted from Ladd's (1996) discussion of data reported by Bruce (1977), in which certain pitch contours in Swedish are shown to display localized interactions between word accent ('acute' vs. 'grave') and phrase accent.

The key point of this section is the following. Valid phonological interpretation of acoustic/articulatory data requires a bridge in the form of a quantitative model that satisfies two criteria: First, it must provide an adequately detailed characterization of the data. Second, the model must have clear links with the phonological level. Whereas the P-P approach can be faulted to primarily pay attention to the second of these criteria by imposing – in the form of 'targets' and 'anchors' – a simplistic, discrete structure on a fundamentally continuous acoustic/articulatory reality, the superpositional framework can be faulted for not having well-developed clear links to the phonological level. It is our hope, of course, that such links can be developed.

3. Linear Alignment Model

The Linear Alignment Model was initially developed to quantify how peak location depends on segmental and durational factors. It then became clear that the model could be reformulated as a superpositional model (van Santen & Hirschberg 1994; van Santen & Möbius 1999). We first describe how it predicts peak location, and then re-characterize it as a superpositional model.

3.1. Linear Alignment model: Predicting the Peak Location

The model states that, in rise-fall patterns, peak location measured from the start of the accented syllable as a function of the durational and segmental composition of a foot is given by:

(2)
$$T_{peak}(D_{onset}, D_{s\text{-}rhyme}, D_{rest}; C_o) = \alpha_{Co} D_{onset} + \beta_{Co} D_{rhyme} + \gamma_{Co} D_{rest} + \mu$$

Here, D_{onset} is the duration of the consonants in the accented syllable onset (excluding any non-syllable-initial sonorants); $D_{s\text{-}rhyme}$ is the duration of the nucleus including any preceding non-syllable-initial sonorants in the onset and sonorants in the coda; D_{rest} is the combined

duration of the remaining deaccented syllables in the foot; α_{Co}, β_{Co}, and γ_{Co} are weight parameters that depend only on the segmental composition of the onset consonants as defined in terms of broad phonetic classes (C_o = voiceless, voiced non-sonorant, or sonorant); and, finally, μ characterizes the delay of the start of the rise relative to the start of the foot. The weight parameters are estimated using standard linear multiple regression, with D_{onset}, $D_{s\text{-}rhyme}$, and D_{rest} as predictor variables and peak location, T_{peak} as dependent variable.

We define a foot in the classical Abercrombie sense as the (left-headed, or trochaic) foot consisting of an accented syllable followed by zero or more unaccented syllables without regard to word boundaries, and as not the Jassem Narrow Rhythm Unit sense (Bonzon Hirst 2004). We surmise that using this type of foot as a structural unit, or, for that matter, using the foot concept for this purpose at all, is likely to be language dependent. Usage of the foot as the structural unit is at this point a working hypothesis requiring more research.

3.2. Estimation of Linear Alignment Model Parameters

The model was applied to simple phrases of the type 'Now I know *word*', with an H* pitch accent on *word* and a low boundary tone, and where pitch accents across different words were perceived as equivalent, with perhaps slight prominence strength variations. 1727 words were recorded from a single speaker. Results were as follows.

First, a correlation of 0.87 was obtained between observed and predicted peak locations. This is a powerful correlation, given the large number of observations (1727) and the small number of parameters [ten: one for the μ parameter and nine for the combinations of onset consonant class C_o and the three parameters α, β, and γ].

Second, variants of the model, specifically variants where the weight parameters were also made to depend on other consonants besides the onset consonant, produced the same correlation.

Third, the estimated values of the weight parameters were largest for voiceless onsets and smallest for sonorant onsets. We note that the split of the foot into the onset, s-rhyme, and the unstressed remainder is fairly arbitrary, and largely dictated – in particular the definition of s-rhyme – by goodness-offit considerations. More broadly, an important question is how to represent the segmental structure in the model; perhaps it should be represented using different chunks, or perhaps one should do away with chunking and represent it as a continuous "sonority curve" based on an

understanding of the dynamics of the larynx and its interactions with superglottal constriction.

Fourth, it was found that the estimated value of μ was essentially zero and its elimination from the model did not significantly affect the 0.87 correlation. This is supported by results in several language showing that the start of the rise coincides with – is 'anchored at' – the start of the accented syllable (Caspers & van Heuven 1993; Arvaniti *et al.* 1998; Ladd *et al.* 1999; Ladd *et al.* 2000; Schepsson *et al.* in press).

Fifth, the estimated values of the parameters α, β, and γ showed the following pattern (compare to parameter values in Figure 2, for the 'peak' anchor point).

(3) $1 > \alpha > \beta > \gamma > 0$

This contradicts various simple models of peak placement, such as: fixed percentage into the syllable (because $\alpha \neq \beta$), fixed percentage into the vowel or s-rhyme (because $1 > \alpha$), and fixed 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 duration(s) of the latter. This adds credence to our assumption that not the accented syllable but the foot is the more appropriate structure unit for understanding intonation.

3.3. Application of the Linear Alignment Model to Other Studies

To give the reader better intuitions about how the model works, we analyze results from a series of studies by Ladd and his colleagues on alignment in English and Dutch (Ladd *et al.* 1999; Ladd *et al.* 2000; Schepman *et al.* in press). These studies are among the few that provide sufficient quantitative detail to make these analyses possible.

We first apply the model to results from a study on effects of speaking rate (slow, normal, fast) on peak placement in prenuclear pitch accents in English (Ladd *et al.* 1999). It is not known whether speaking rate differentially affects phonemes in onsets, s-rhymes, and unaccented syllables. As a first order of approximation, we assume that these effects are proportionally the same (i.e., same %) across phonemes and phoneme locations. If, then, based on the previous paragraph, we drop μ in Eq. 2 and include the speaking rate parameter R as a proportionality constant, we can write:

$$T_{peak}(D_{onset}, D_{s\text{-}rhyme}, D_{rest}; R) = \alpha R D_{onset} + \beta R D_{rhyme} + \gamma R D_{rest}$$

In the ratio of $T_{peak}(D_{onset}, D_{s\text{-}rhyme}, D_{rest}; R)$ and the duration of the sum of the onset and the s-rhyme, i.e.,

$$\frac{T_{peak}(D_{onset}, D_{s-rhyme}, D_{rest}; R)}{R D_{onset} + R D_{s-rhyme}}$$

the rate parameter R cancels. Thus, the ratio is independent of speaking rate; the ratio does depend on the values of the parameters α, β, and γ, which may be speaker dependent. Analysis shows that, averaged over speakers, the values of the ratio for the three speaking rates are 0.948, 0.958, and 1.040. Across speakers the ranges of these three ratios are also reasonably narrow and clearly speaker dependent (1.05-1.13, 0.79-1.03, 0.77-0.85, 0.93-1.08, 0.85- 0.93, and 1.09-1.37; the last speaker had an exceedingly slow slowest speaking rate and may be considered an outlier). We tentatively conclude that the ratio remains fairly constant across speaking rates within speakers, as predicted by the model.

A second relevant study concerns the effects of vowel duration on peak location in Dutch (Ladd *et al.* (2000), Experiment 1). In this study, the peak was positioned slightly before the end of the accented vowel for long vowels (12 ms), and after the end of the vowel for short vowels (25 ms). Equivalently, since the durations of the long and short vowels were 133 and 77 ms, the peaks were positioned at 121 and 102 ms after the start of the accented (long or short) vowel. This 121 ms - 102 ms, or 19 ms, effect of vowel duration on peak location is predicted by Eq. 2, assuming that D_{onset} and D_{rest} were not significantly affected by the long- and short-vowel contexts. Under this assumption, the difference in peak location is given by

$$\beta\,(133 + \textit{sonorant coda duration}) - \beta\,(77 + \textit{sonorant coda duration})$$

or 56β ms. Since the model states that $0 < \beta < 1$, it correctly predicts that the effect of vowel duration on peak location, or 19 ms, should be between 0 and 56 ms. Using an estimated value for β of 0.5 from our studies, the predicted effect of vowel duration on peak location is 28 ms. The 19 ms effect corresponds to an β value of 0.3 averaged over the 6 informants, with individual values ranging between 0.2 and 0.5. Several factors could be responsible for the smaller size of β compared to our studies, including inaccuracy of the assumption of equal

durations of the remaining phonemes, usage of prenuclear vs. nuclear pitch accents, or the language differences (Dutch vs. American English).

A study by Schepman, Lickley, and Ladd (in press) is similar to the experiment discussed in the preceding paragraph in that it manipulated vowel length. However, the study also manipulated stress clash and analyzed nuclear instead of prenuclear pitch accents. The effect on peak location of vowel duration, averaged over stress clash condition, was 22 ms, representing 35% of the 62 ms difference in vowel plus onset duration (221 vs. 159 ms) and yielding a value of β of 0.35.

The implications for the model of the results obtained in Experiment 2 in Ladd *et al.* (2000), however, are more ambiguous. The experiment was the same as Experiment 1, but instead of contrasting phonetically long and short vowels, the study contrasted phonologically long and short vowels that have the same duration (/i:/ vs. /I/). As measured from the accented vowel onset, peak locations were located at 85.1 ms and 94.3 ms, respectively, with vowel durations of 64.5 and 62.8 ms. Clearly, the model cannot explain this 94.2 ms - 85.1 ms, or 9.2 ms, effect on peak location on the basis of the difference in vowel duration; by the same logic as above, it would predict a slight (< 2 ms) effect in the opposite direction.

There are two caveats, however. First, the onset consonants were not quite matched between the two conditions, with more voiceless onset clusters (e.g., /sp/) in the /I/ condition, thereby causing longer durations of Donset in the /I/ condition and hence potentially explaining the 9.2 ms effect, taking into account that alignment parameters have the larger values for voiceless onsets than for voiced obstruent or sonorant onsets (van Santen & Hirschberg 1994). Second, and more important, the authors noted that in Dutch postvocalic consonants can be viewed as belonging to the onset of the next syllable when they follow phonologically long vowels, but as (at least partially) belonging to the coda when they follow phonologically short vowels. If so, then $D_{s\text{-}rhyme}$ is in fact longer in the short vowel condition due to (partial) migration of the (in the experiment, invariably sonorant) postvocalic consonants to the sonorant rhyme, and D_{rest} is correspondingly shorter. Since the weight parameter applied to the sonorant rhyme (β) is larger than the weight parameter applied to the post-rhyme part of the foot (γ), the observed results could be due to this re-syllabication as well. But without further experiments, these two caveats require further confirmation.

We draw the following tentative conclusions from the illustrative examples. First, the simple model is able to account for some fairly

complex data patterns. Second, this account can lead to different interpretations than accounts based on the P-P method of alignment measurement. Specifically, where the P-P method may see differences in pitch alignment, our model may only see effects of segmental-temporal structure on the time course of the pitch curve. Third, instead of casting the results in terms of anchoring of pitch points at segmental points, the model views the location of pitch points as the complex resultant of the durations and segmental makeup of parts of the foot. Although it refers to points in the segmental stream to delineate these parts, it does not require a pitch point to be anchored at any of these points – any such apparent anchoring is accidental and carries no phonological or perceptual implications. Fourth, the model suggests a method for measuring alignment that is substantially different from currently used methods. The model implies that one can draw inferences about the effects of some factor (e.g., phonological vowel length) only by fitting the model and comparing the estimated weight parameter values between the conditions defined by this factor. Much more about this will be said in Section 4.

3.4. Linear Alignment Model: Superpositional Characterization

All results reported so far were based on peak location, for reasons of convenience and tradition. Thus, in our effort to question the P-P

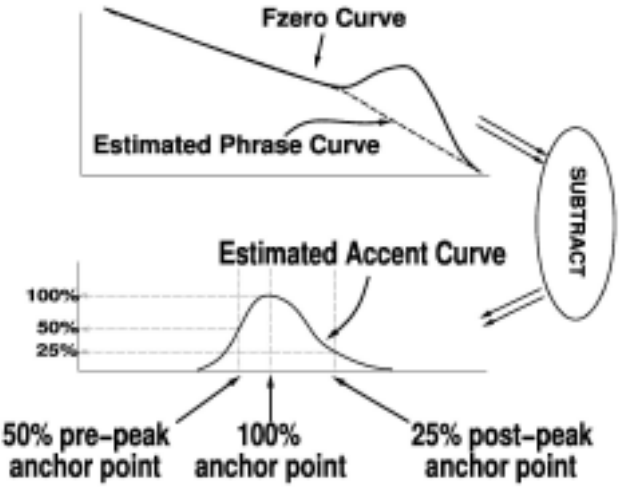

Fig. 1. Estimation of accent curves and anchor points. The top panel shows the generation of the phrase curve; the bottom panel shows the shape of the estimated accent curves and the computation of anchor points.

approach, we are only half-way, having done away with the reliance on segmental points but not yet with reliance on pitch points. Towards this end, we re-cast the model for peak location (i.e., Eq. 2) using the superpositional approach. Because of the extreme simplicity of the pitch contours studies in (van Santen & Hirschberg 1994) resulting from utterances that contained a single pitch accent in a carrier phrase, it is easy to draw a line from the start of the pitch accent to the end of the phrase (Fig. 1).

Subtraction of this line from the pitch curve produces a curve that rises from zero to a peak value, and then returns to zero (Figure 1, bottom panel).

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. For example, the anchor point that is located to the left of the peak anchor point and that has an anchor value of 50% 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, pointof- steepest ascent, and the like.

The generalization of Equation 2 to accommodate the anchor point concept is obvious (*A* refers to the *A*-th anchor point):

$$T_A(D_{onset}, D_{rhyme}, D_{rest}) = \alpha_A D_{onset} + \beta_A D_{rhyme} + \gamma_A D_{rest} \quad (4)$$

We call the ensemble of parameters α_A, β_A, and γ_A alignment parameters. Figure 2 shows their values for polysyllabic contexts (i.e., the accented syllable was followed by at least one unaccented syllable). Again, as in the case of simpler model in Eq. (1), these parameters are estimated using standard linear multiple regression.

We note that these anchor points comprise a discrete approximation to a continuous reality – they are sampling points in the same sense as samples in analog-to-digital conversion of speech. This underlying continuity is emphasized in Figure 2 by connecting the points with smooth curves. As in analog-to-digital conversion of the speech signal, the particular selection of sampling points is not critical provided that these points are sufficiently dense. Virtually the

same rise-fall pattern (i.e., $T_A(D_{onset}, D_{rhyme}, D_{rest})$, can be characterized by any number of anchor point selections. Thus, these anchor points have no special phonological status, and certainly should not be thought of as (together with their respective values of $T_A(D_{onset}, D_{rhyme}, D_{rest})$ as 'targets' (see also Section 4.3).

We also want to note that usage of these sampling points does not constitute a theoretical difference with the Fujisaki model, whereby the latter would be a continuous model and the current model is a discrete model. Both models are continuous, and the discretization only enters in the estimation process of the Linear Alignment Model.

In what follows, we describe how Eq. 4 can be used for the generation of accent curves within a superpositional framework.

This simple procedure for defining points on the time axis (i.e., T_A) is tacitly assuming the superpositional model, whereby the subtracted line plays the role of (local estimate of) phrase curve and where the curve that remains after subtraction is the sum of the accent curve and the segmental perturbation curve (see below). In this section we formalize this.

The Linear Alignment Model uses three curve classes: Phrase Curves, Segmental Influence Curves, and Accent Curves. In some versions, such as for Japanese (Venditti & van Santen 2000), further curve classes are added such as "UA curves" (defined as 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 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. Interestingly, this account differs from the traditional account in which the curve is seen as containing a flat, straight plateau that ends at the fall in the accented syllable. Rather, we see a small but distinctive inflection point right at the start of the accented syllable that can be accounted for by the underlying accent curve and that contradicts the straightness of this region.

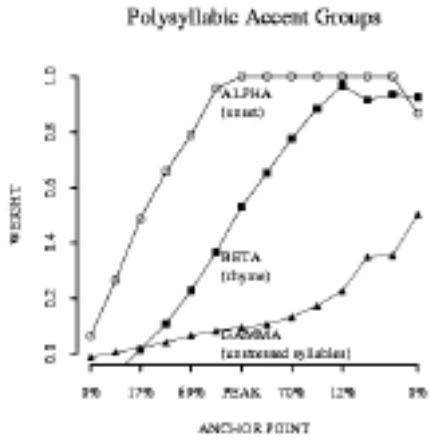

Fig. 2. Alignment parameters.

3.4.1. Phrase Curves

Whereas the Fujisaki model makes strong assumptions about the phrase curve, in the 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 (van Santen *et al.* 1999), 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.

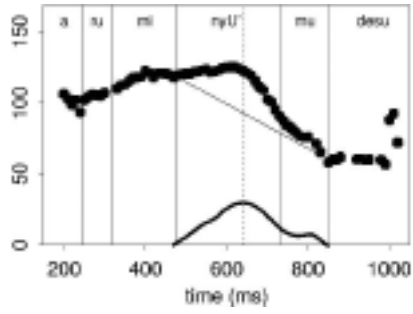

Fig. 3. Analysis of a UA-group F_0 contour. The bottom curve is the accent curve that accounts for the late fall in the accented mora when a local UA curve (thin line) is subtracted from the F0 curve. From: Venditti & van Santen, "Japanese Intonation Synthesis using Superposition and Linear alignment", Proc. ICSLP 2000.

3.4.2. 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.

3.4.3. Accent Curves

In most speech synthesis 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 (van Santen & Hirschberg 1994), where we focused on syllables carrying nuclear pitch accents and where these syllables where followed by varying numbers of unaccented syllables.

In the Linear Alignment Model, accent curves are generated in a way that differs fundamentally from the Fujisaki model. Specifically, we make use of Eq. 4 to generate accent curves 'from templates via parameterized time warp functions'.

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_A, β_A, γ_A, for all anchor points, A. Given a rendition of trochaic foot with durations of D_{onset}, $D_{s\text{-}rhyme}$ (or D_{rhyme} and D_{rest}), the A-th anchor point is located on the time axis as indicated by Eq. 4, 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 their segmental/ durational structures are different. If some factor affects the segmental/duration structure of a foot, this may cause changes in alignment as measured by the P-P approach; however, these changes may not be changes in alignment in terms of the Linear Alignment Model.

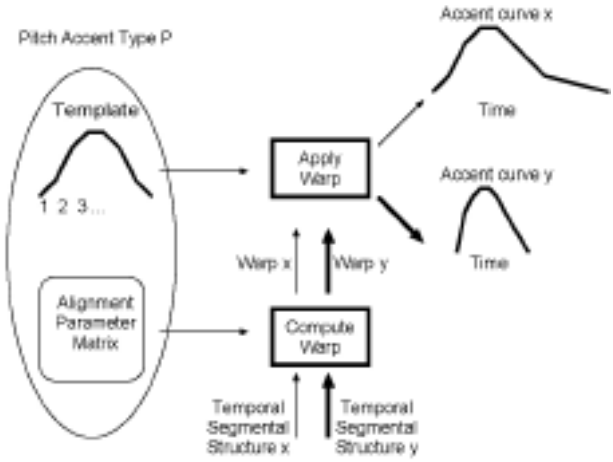

Fig. 4. Flow diagram of accent curve generation. Temporal pattern *x, y* are combined with a pitch-accentspecific alignment parameter matrix to form time warps, warp *x*, warp *y*, that are applied to a pitchaccentspecific template to generate accent curves.

4. Implications of the Linear Alignment Model

We now address the implications of the Linear Alignment Model for the measurement of alignment. First, this section briefly touches on additional questions, specifically concerning phonetic invariance and the concept of 'target'

4.1. Perceptual effects of small displacements

How does the model account for the changes in phonological category associated with small displacements (D'Imperio & House 1997)? 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 segmental/temporal structures on the one hand to accent curves on the other hand. In other words, they define how accent curves and segmental/temporal structures are coordinated. According to the Linear Alignment Model, the change in perceived phonological category associated with small displacements is due to the fact that these curves 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 segmental/temporal structure the same. Thus, even though the shape, and hence presumably the underlying templates, were the same, the alignment parameters cannot also have been the same. Of course, in van Santen and Hirschberg's production studies, always segmental/temporal structure varied but pitch accent class was the same.

4.2. Practical advantages of avoiding peaks

The following advantages exist 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.

4.3. The concept of target

Pitch contours are commonly described in terms of movement between 'tonal targets'. Within the P-P approach, such targets are defined as points in Time × Frequency space (or in a relative frequency space, as a fractional quantity relative to 'reference' and 'top' lines (Pierchumbert 1980; 1981). But also proposals have been made for dynamic targets. For example, Xu (2001, 2002) models the surface pitch contour in terms of (unidirectional) pitch movements between pitch "levels" (i.e., horizontal line segments in Time × Frequency space) or pitch "directions" (i.e., non-horizontal line segments in Time × Frequency space). In the IPO approach ('t Hart *et*

al. 1990), targets consist of line segments that are characterized in terms of slope and temporal extent; the pitch curve consists of a smoothed, connected sequence of these line segments. In the Tilt Model (Taylor 2000), the F_0 contour is modeled by a one- or two-line segment contour inside the accented syllable and by smooth interpolation between accented syllables.

What does the Linear Alignment Model have to say of what speakers are aiming at? The model does not in any sense use the target concept. According to the model, the speaker aims at producing a particular accent curve shape that is appropriately aligned with the segmental/temporal structure of the segment stream; the shape and the alignment are dictated by the accent template and the alignment parameter matrix. The frequency/time coordinates of the surface pitch contour are the complex resultant of the phrase curve, the accent curve, and segmental perturbation, and are not in any direct way "aimed at" by the speaker. Thus, according to the model, the speaker does not aim at tonal targets in absolute or relative Time ? Frequency space, but aims at (uni- or bi-directional) accent curves that may have a very oblique relationship with the surface pitch contour (as exemplified, e.g., by the "invisible" accent curves in (Venditti & van Santen 2000).

A further argument against tonal targets as a valid construct in non-tone languages comes from contrasting them with phoneme targets. Phoneme targets can be approximately characterized either as single points in some articulatory/acoustic space, or (in diphthongs, stops, affricates) as a – smooth or non-smooth – movement between a succession of such points. If one were to listen to a speech signal synthesized from these points by excising a spectral vector from the center of a segment or sub-segment and using signal processing to extend it in time, the thus generated speech fragment would be identifiable as being associated with the segment or sub-segment in question. It thus makes sense to loosely describe speech as a sequence of (possibly multi-tiered) movements from one phonemic target to the next (Kain & van Santen 2000).

However, when we excise an interval of speech surrounding the peak portion of a rise-fall F_0 pattern, the excised speech obviously does not in any meaningful way resemble that rise-fall pattern. We thus speculate that, at least in nontone languages, (uni- or multi-directional) pitch movement may typically be the minimal element of intonation meaning.

4.4. Measurement of Alignment

The paper thus far has two implications for measurement, one is that measurement is inevitably model based and, secondly, that measurement of alignment may be more complex than can be captured with the P-P approach to measurement.

We now describe how the Linear Alignment Model is used for measurement of alignment. The measurement procedure can be generalized from the process described in Section 5.1 (Figure 1, and Eq. 4). For illustrative purposes, we assume that the research question of interest is whether pitch alignment in context C_1 is later than in context C_2 (e.g., the contexts may be declaratives vs. interrogatives in Neapolitan Italian (D'Imperio & House 1997); the expression of *knowing*, 'recognizing', vs. 'surprise' (Kohler 2004); or speaking rate (Ladd *et al.* 1999).

Recordings need to be made of multiple instances of each context; these instances should also vary substantially in segmental/durational structure. Their segmental structures should be matched to the greatest degree possible across contexts, but this is not essential. It is assumed that the different contexts are associated with similar-looking pitch accents (e.g., rise-fall only, or rise-only).

(1) Determine the durations of the appropriate sub-intervals (e.g., onset, srhyme, and remainder of foot) of the structural unit associated with pitch accent (e.g., foot).

(2) Compute the local phrase and segmental perturbation curves, and subtract from observed pitch curves to produce estimated accent curves and anchor point locations. We note that this problem has not been solved in general; in our earlier studies, estimation of phrase curves was simple because of the simplicity of the curves studies (van Santen & Hirschberg 1994).

(3) Two options:
(a) For each context separately, estimate alignment parameters.
(b) Combine the two contexts, estimate (shared) alignment parameters using multiple regression and compare predicted vs. observed anchor point locations for each context.

Alignment can be said to be later in C_1 than in C_2 either when (a) the corresponding alignment parameters have smaller values for C_1 than for C_2, or, alternatively, when (b) observed anchor point

locations in C_1 are systematically later than their predicted locations and when observed anchor point locations in C_2 are systematically earlier than their predicted locations, where both predictions are based on shared alignment parameter estimates.

This procedure measures anchor point locations for arbitrary accent curve shapes, and does so in a way that fully takes into account any effects of context on the segmental/segmental structure of the segment stream. The procedure also provides a detailed picture of alignment that goes beyond 'early' and 'late'. For example, if the alignment parameter curves cross, then this may mean that in one context the rise portion is faster while in the other context the fall portion is faster. Also, the procedure provides information about the relative impacts on pitch timing of the respective sub-intervals of the structural unit.

5. Applying the Linear Alignment Model

In the preceding sections, we have explained the Linear Alignment Model and how it can be used as a theoretical framework for the study of alignment and as a method for measuring alignment. However, as stated before, this is a far cry from having the ability to actually use the model for the measurement of alignment for arbitrary pitch contours. A fundamental obstacle is that because, unlike the Fujisaki model, the Linear Alignment Model does not make strong assumptions about the shapes of its curve classes, we have a difficult situation in which we need to develop minimalist assumptions about some aspects of the model in order to estimate other aspects of the model.

The research program on which we report next has as its goal that of filling in the many undeveloped aspects of the Linear Alignment Model. This research is conducted in the context of a project on highly expressive read speech, specifically, story telling to small children. Compared to the reading style studied most often, news reading, this style is characterized by a much larger pitch range and the presence of a much greater variety of local pitch movement shapes (Klabbers & van Santen 2002). Also, prosodic cues other than pitch, loudness, and timing are used, or are used to greater effect, in particular phonation mode and oral aperture – not to speak of facial expression. The present discussion, however, is confined to pitch.

The key intermediate goals are these, and include both broad and narrow technical goals. First, what is the collection of local pitch

curves in this reading style and how do we tag them? Second, how can we decompose pitch curves into component curves without making strong assumptions except for additivity? Third, how can we fill in 'gaps' in pitch contours that occur in non-sonorant regions or in sonorant regions, due to post-obstruental effects?

5.1. Foot-based Tagging Schemes

Klabbers and van Santen (2002) compared multiple schemes for tagging syllables. One group of schemes was foot based and tagged syllables in terms of the location of the syllable in the foot (number of syllables to the left and to the right in the foot), and the location of the foot in the utterance (phrase-medial, phrase-final, and utterance-final). The other tagging schemes were syllable based and did not directly or indirectly refer to the feet. The basic research question was: Which scheme explains the most variance of pitch curves with the smallest number of categorical distinctions?

For variance, we used a measure of 'within-class pitch contour variance' that reflected differences in direction of pitch movement as well as excursion magnitude, while being relative insensitive to underlying phrase curve differences.

The following were the main results. First, the foot-based tagging scheme produced a classification that was more homogenous – had less within-class variance – than other classifications. However, the within-class contour variance was substantial for all schemes. Second, comparison of two foot-based schemes one with vs. and the other without reference to location of the foot in the utterance, showed that foot location matters. Third, the data contradicted a key assumption of the foot based scheme, which is that for foot-initial syllables (i.e., the accented head of a syllable) pitch should be unaffected by whether it is preceded by zero, one, or more unaccented syllables. Specifically, the data showed that stress clash (i.e., this number is zero) had an effect, but that there was no difference between the number being one vs. more than one. This implies that stress clash must be included in the tagging scheme. In fact, it may be used to re-define feet (e.g., using iambic reversal, i.e., moving the pitch accent of the left word earlier in that word to avoid stress clash as in tennesSEE is a red state vs. TENnessee VALLey).

These and other results, while showing the empirical value of the foot, also show that pitch contours in high-expressivity speech have sources of variation that cannot be captured by tagging schemes based on standard linguistic markers alone. In this type of speech,

many other descriptors, some affective and others semantic, must be added.

5.2. Clustering of Accents

In order to further explore foot tagging schemes and the shapes of accent curves associated with foot classes, Klabbers and van Santen (2004) conducted a clustering study. The data analyzed consisted of fragments of pitch curves associated with feet. The same distance measure was used as for computing pitch contour shape variance in the previous study (Klabbers & van Santen 2002), except that the shapes were time warped to generate optimal mutual alignment between curve fragments. This time warping process was done as follows. First, for those feet containing a standard rise-fall pitch movement, the average peak location and foot duration were computed for each combination of foot length and foot position (phrase-medial, phrase-final, and utterance-final), and their ratios were computed. For example, for monosyllabic feet, these ratios were 54, 41, and 40% for the three respective foot positions. We assumed that these percentages could be applied to any type of pitch contour, not only rise-fall contours, to normalize its time course by using the predicted ratio (predicted on the basis of foot length and foot position) to locate a central anchor point, and time-normalizing the portions of the curve preceding and following this point. As in the Linear Alignment Model, it is thus assumed that, even for curves that have no peaks, this normalization accurately reflects the non-linear effects on the time course of the pitch curve of the durations of the syllables that make up a foot. Still in other words, we assume that, to a first order of approximation, the alignment parameters are the same for these different shapes in this study.

Key results were the following. Six clusters were discovered. One of these clusters exhibited the typical up-down movement where the peak is associated with the head of the foot. However, other clusters were more surprising. The most important one was that two feet (most frequently occurring at the end of a minor or major phrase) can be connected by what seems to be a different type of phrase curve consisting of a rise on the first foot and a fall on the second foot. Depending on the length of the penultimate foot there may be a plateau in between these two movements. A second important observation is that the continuation rise which was always assumed to be present at minor phrase boundaries was observed in fewer than 10% of feet occurring at the minor phrase boundary in this corpus.

Some of these occurred in reading list-type sentences ([e.g., blueberries, strawberries, and bananas. (Mishra *et al.* 2003). But not every item in the list showed a continuation rise.

Similar to the previous study, also this study showed that efforts need to be made to enrich the tagging scheme because the six clusters cut across the classes of the tagging scheme proposed in (Klabbers & van Santen 2002). A possible implication of the rise-only and fall-only clusters is that the Linear Alignment model must be equipped with phrase curves that contain bulges or even plateaus.

5.3. Automated decomposition of pitch curves into component curves

The studies by Klabbers and van Santen (2002, 2004) could have benefited substantially from having tools that can decompose arbitrary pitch curves into component curves. This is an extremely hard problem, because in order to estimate these curves we may need to make strong assumptions about their shapes, which is exactly what we want to avoid. Nevertheless, preliminary results using a 'wavelet based approach' seem promising (van Santen *et al.* 2004).

The additivity of the Linear Alignment Model immediately suggests using some sort of Fourier transform to extract the component curves. And indeed, this method is used as an important component in a Fujisaki parameter estimation system developed by Mixdorff (2002), and was proposed earlier by Sakurai and Hirose (1996). Our experiments showed, however, that this method provided undesirable results. Depending on the frequency cutoff, the result was either a smooth but poorly fitting phrase contour or an irregularly shaped phrase contour that contained the negative lobes of frequency components that carried the accent curves.

Wavelets share with the Fourier transform the features of linearity and frequencyspecificity. The key difference is the temporal 'locality' of the wavelet transform. That is, like the Fourier transform, the wavelet transform of an input signal (here, the F_0 curve) represents the signal as a weighted sum of basis functions. However, these basis functions consist of dilations (time scaling) and translations (moving along the time axis) of a localized (at τ_0) wavelet function $\Psi(t - \tau_0)$. Thus, unlike the Fourier transform, the behavior of same-sized (or, equivalently, same-frequency) basis functions at different locations is independent. This basic feature, so we reasoned, might be critical in addressing the irregularly shaped phrase contour problem encountered with the Fourier transform.

Application of the wavelet transform, eliminating certain transform components (much like band-pass filtering), and computing the inverse of the thus filtered transform produced promising results. The method was applied to pitch curves generated by the Fujisaki model and the Bell Labs intonation model, which gave us the opportunity to compare the extracted phrase curves with the known curves in these models. Results show that accurate estimates of phrase curves can be obtained for pitch curves generated by these models, even though they have differently shaped phrase and accent curves. Importantly, the choice of transform components to eliminate was the same for both models – there was no model-specific fine-tuning.

This method, while certainly inadequate for highly complex pitch contours, nevertheless will be a useful component of a more complex system in which also knowledge of foot boundaries and other tags are known.

5.4. Filling-in Gaps using Prosody Copy

A significant obstacle for any type of analysis of pitch contours, whether quantitative or qualitative, is the fact that only some parts of these contours can be trusted. Other parts are either absent (in voiceless regions), likely to contain pitch tracking errors (in voiced obstruents), or affected by preceding obstruents. A principled way is needed to fill in these gaps.

The traditional method for dealing with these problems consists of smoothing and linear interpolation in gaps. A fundamental problem with this method is that, in terms of the Linear Alignment Model, it does not eliminate effects due to the segmental perturbation curve. In an example mentioned earlier, in the word *pit*, segmental perturbations that follow the prevocalic obstruent may extend 50-100 ms into the vowel. These perturbations can either create a spurious peak right after the vowel onset, or, after smoothing, a flat region extending from the vowel onset to the true peak. In either case, the resulting maximum does not coincide with the 'true' peak. The key reason for this is that segmental perturbations cannot be characterized as locally random (e.g., jitter), but instead should be viewed as systematic effects that cannot be eliminated by smoothing or interpolation.

The preferred method for filling in gaps caused by any of these factors should be based on a clear understanding of the shapes of the component curves. Mishra and van Santen are developing a prosody copy method, which works as follows. Consider an appropriately

labeled speech corpus. For a given 'target foot' that contains gaps of any type, consider the set of all 'similar all-sonorant feet' ('SAS feet') in the corpus, i.e., feet that are similar to the target foot in the sense that they have the same basic shape (as determined by an appropriately defined shape similarity measure) and the same foot tag (e.g., using the Klabbers and van Santen (2002) tags). Based on the tag, we compute a time warp in the same way as in (Klabbers & van Santen 2004), and use this warp to perform a temporal normalization of both the target foot and the SAS feet. Each of these feet is then characterized by two vectors of equal length (n). One of these contains the pitch values, $\vec{F}_0$ (which may include missing values). The other contains weights, $\vec{W}$. These weights are computed using a 'sonority measure' (e.g., the product of a voicing flag and amplitude), further refined by setting weights equal to zero for suspicious areas such as the first 50 ms of a post-obstruental vowel or small areas surrounding vowel-nasal or nasal-vowel boundaries.

If the time warp operates as intended, then – missing values permitting – the shapes of the pitch vectors should be highly similar to each other and to the target contour. For example, if the shape is single-peaked, then the time warp should produce pitch vectors that all contain the peak values in the same location of the vector.

The final step consists of using weighted linear regression to fit the following model, where i is a SAS foot, and n is the vector containing the numbers 1 through n, and $\vec{1}$ the vector consisting of n 1s:

$$\vec{F}_0^{[target]} \approx \alpha \vec{F}_0^{[i]} + \beta\vec{1} + \gamma\vec{n} \tag{5}$$

The weights used in the regression analysis are given by the direct product of $W^{[target]}$ and $W^{[i]}$.

If (i) the phrase curves are locally linear, (ii) the (hidden, not explicitly estimated) accent curves of the target and SAS feet are proportional to each other (i.e., only differing in height), and (iii) the weights in $W^{[target]}$ are zero for any regions containing segmental perturbations, then it can be shown that this approximation provides an unbiased estimate of the sum of the local phrase curve and the accent curve. Preliminary results show that this works reasonably well, provided that a speech corpus contains adequate SAS feet for a given target foot.

6. 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 on the core issues concerning the phonetics of alignment: What is aligned, and what stays the same when segmental/temporal structure varies? The hypothesis that intonationalphonological classes are associated with combinations of templates and alignment parameter matrices, while currently highly speculative, exemplifies a deeper principle according to which we must search for fairly complex forms of invariance at the articulatory/acoustic level, and not be satisfied with simple surface features such as pitch peaks. Of course, much remains to be done at the theoretical level. For example, the dependency of the alignment parameters on the phonemic class of the onset consonants (C_0) shows that the model is in need of a deeper level of explanation in terms of articulatory principles, such as the effects of prevocalic consonant class (voiceless consonants, voiced non-sonorants, vs. sonorants) on laryngeal dynamics.

The reported studies conducted in our laboratory also illustrate that much remains to be done at the practical level to make the Linear Alignment Model serve the stated purpose of being useful as a tool for measuring alignment. The core obstacle is our hesitation to make strong, but arbitrary, assumptions without a critical mass of converging evidence; the willingness to make such assumptions is both a strength and a weakness of the Fujisaki model.

Address of the Authors:

Center for Spoken Language Understanding, OGI School of Science & Engineering, Oregon Health & Science University

Note

* This material is based on work supported by the National Science Foundation under Grants No. 0205731 ("ITR: Prosody Generation for Child Oriented Speech Synthesis"), jointly with Alan Black and Richard Sproat; 0313383 ("ITR: Objective Methods for Predicting and Optimizing Synthetic Speech Quality"); and 0082718 ("ITR: Modeling Degree of Articulation for Speech Synthesis"). We thank Bob Ladd and Yi Xu for insightful comments on an earlier draft of this paper.

References

ARVANITI Amalia, D. Robert LADD & Inexe MENNEN 1998. Stability of tonal alignment: the case of greek prenuclear accents. *J. Phonetics* 26. 3-25.

BOUZON Caroline & Daniel HIRST 2004. Isochrony and prosodic structure in British English. In *Proceeding Speech Prosody 2004*. Nara. Japan.

BRUCE Gösta 1977. *Swedish word accents in sentence perspective*. No XII in Travaux de l'Institut de Phonétique. Lund: Gleerup.

CASPERS Joanneke & Vincent VAN HEUVEN 1993. Effects of time pressure on the phonetic realization of the dutch accent-lending pitch rise and fall. *Phonetica* 50. 161-171.

D'IMPERIO Mariapaola 2002. Language-specific and universal contraints on tonal alignment: the nature of targets and 'anchors'. In *Proceedings of Speech Prosody*. Aix-en-Provence.

D'IMPERIO Mariapaola & David HOUSE 1997. Perception of questions and statements in Neapolitan Italian. In *Proceedings of the Fifth European Conference on Speech Communication and Technology*. Rhodes. September.

FUJISAKI Hyroya 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In MACNEILAGE Peter F. (ed.). *The production of speech*. New York: Springer. 39-55.

FUJISAKI Hyroya 1988. Note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In FUJIMURA Osamu (ed.). *Vocal Physiology, Voice Production, Mechanisms and Functions*. Raven Press. 347-355.

FUJISAKI Hyroya 1995. Physiological and physical mechanisms for tone, accent and intonation. In *Proceeding of the XXIII World Congress of the International Association of Logopedics and Phoniatrics*. 156-159.

KAIN Alexander & Jan VAN SANTEN 2002. Compression of acoustic inventories using asynchronous interpolation. In *Workshop on Speech Synthesis*. Santa Monica. California. IEEE.

KLABBERS Esther & Jan VAN SANTEN 2002. Prosodic factors for predicting local pitch shape. In *Workshop on Speech Synthesis*. Santa Monica. California. IEEE.

KLABBERS Esther & Jan VAN SANTEN 2004. Clustering of foot-based pitch contours in expressive speech. In *5th ISCA Speech Synthesis Workshop*. Pittsburgh. PA. IEEE.

KOHLER Klaus 2004. Prosody revisited: function, time, and the listener in intonational phonology. In *Proceeding Speech Prosody 2004*. Nara. Japan.

LADD D. Robert 1983. Phonological features of intonational meaning. *Language* 59. 721-759.

LADD D. Robert 1996. *Intonational phonology*. Cambridge: Cambridge University Press.

LADD D. Robert, Dan FAULKNER, Hanneke FAULKNER & Astrid SCHEPMAN 1999. Constant segmental anchoring of F_0 movements under changes in speech rate. *Journal of the Acoustical Society of America* 106. 1543-1554.

LADD D. Robert, Ineke MENNEN & Astrid SCHEPMAN 2000. Phonological conditioning of peak alignment in rising pitch accents in dutch. *Journal of the Acoustical Society of America* 107. 2685-2696.

MISHRA Taniya, Esther KLABBERS & Jan VAN SANTEN 2003. Detection of list-type sen-24 tences". In *Proceedings of Eurospeech*. Geneva. Switzerland. September. 2477-2480.

MIXDORFF Hansjorg 2002. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proceedings of the 26th International Conference on Acoustics, Speech and Signal rocessing (ICASSP 2002)*, vol. 3. 1281-1284.

PIERREHUMBERT Janet 1980. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, September 1980. Distributed by the Indiana University Linguistics Club.

PIERREHUMBERT Janet 1981. Synthesising intonation. *Journal of the Acoustical Society of America* 70, 4. 985-995.

PRIETO Pilar, Jan VAN SANTEN & Julia HIRSCHBERG 1994. Patterns of F_0 peak placement in Mexican Spanish. In *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*. New Paltz. NY. USA. ESCA/AAAI/IEEE. 33-36.

SAKURAI Atsuhiro & Keikichi Hirose 1996. Detection of phrase boundaries in japanese by lowpass filtering of fundamental frequency contours. In *Proceedings ICSLP*. Philadelphia. 817-820.

SCHEPMAN Astrid, Robin LICKLEY & D. Robert LADD. Effects of vowel length and right context on the alignment of dutch nuclear accents. *Journal of Phonetics*. In press.

'T HART Johan, Renè COLLIER & Antoine COHEN 1990. *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press.

TAYLOR Paul 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America* 107, 3. 1697-1714.

VAN SANTEN Jan & Julia HIRSCHBERG 1994. Segmental effects on timing and height of pitch contours. In *Proceedings ICSLP '94*. 719-722.

VAN SANTEN Jan, Taniya MISHRA & Esther KLABBERS. Estimating phrase curves in the general superpositional intonation model. In *5th ISCA Speech Synthesis Workshop.* Pittsburgh. PA. IEEE.

VAN SANTEN Jan & Bernd MÖBIUS 1999. A model of fundamental frequency contour alignment. In BOTINIS Antonis (ed.). *Intonation: Analysis, Modelling and Technology*. Cambridge University Press. In press.

VAN SANTEN Jan, Chilin SHIH & Bernd MÖBIUS 1997. Intonation. In *Multilingual Textto-Speech Synthesis: The Bell Labs Approach*. R. Sproat, Ed. Kluwer. Boston. MA. ch. 6. 141-189.

VAN SANTEN Jan, Chilin SHIH & Bernd MÖBIUS 1997. Intonation. In *Multilingual Textto-Speech Synthesis*. R. Sproat, Ed. Kluwer. Dordrecht: the Netherlands.

VENDITTI Jennifer & Jan VAN SANTEN 2000. Japanese intonation synthesis using superposition and linear alignment models. In *Proceedings ICSLP*. Beijing. China.

XU Yi 2002. Articulatory constraints and tonal alignment. In *Proceedings of Speech Prosody 2002*. Aix-en-Provence.
XU Yi & Q. Emily WANG 2001. Pitch targets and their realization: Evidence from mandarin chinese. *Speech Communication* 33. 319-337.