MODELING PITCH ACCENT CURVES

Jan van Santen and *Bernd Möbius*

Lucent Technologies – Bell Labs, 600 Mountain Avenue, Murray Hill, NJ 07974, USA
{jphvs,bmo}@bell-labs.com

ABSTRACT

Segmental factors can cause large temporal changes in local pitch contours associated with accented syllables ("accent curves"), but these changes are often not phonologically or perceptually significant. Yet, other factors can cause temporal changes that are smaller but nevertheless significant. We propose a model according to which accent curves are (phonologically, perceptually) equivalent when they are generated from the same accent curve template using a shared family of time warp functions. The model is shown to provide a detailed and accurate account of alignment of accent curves over a wide range of segmental configurations.

1. Introduction

Local pitch contours belonging to the same perceptual/phonological class vary significantly as a result of the structure (i.e., the segments and their durations) of the syllables they are associated with. For example, in simple nuclear rise-fall pitch accents in declaratives [6], we found that peak location (measured from stressed syllable start) systematically varied between 150 and 300 ms as a function of the durations of the associated segments. Yet, there are temporal changes in local pitch contours that are phonologically significant (e.g., [2]), even though their magnitudes do not appear to be larger than changes due to segmental effects.

This paper addresses the following question: *What is invariant about pitch contours belonging to the same class?* We propose a model according to which curves in the same class are generated from a common template using the same family of time warp functions. Classes differ either by having different templates or different time warp function families. The model predicts in detail the alignment of an accent curve with the sequence of segments it is associated with.

2. Accent Curve Alignment: Data

To keep this section as empirical and theory-free as possible, the word "accent curve" is used very loosely in the sense of a local pitch excursion that corresponds to an accented syllable, not in the specific sense of the Fujisaki model [1]. The term "accent group" (or "stress group") refers to a sequence of syllables of which only the first is accented. Finally, "accent group structure" refers to the segments in an accent group ("segmental structure") with associated durations. Thus, renditions of the same accent group almost always have different structures (because their timing is unlikely to be identical), but by definition they have the same segmental structure.

Our data base is an extension of the speech corpus described in a previous paper [6], and consists of speech recorded from a female speaker who produced carrier phrase utterances in which one or two words were systematically varied. The non-varying parts of the utterances contained no pitch accents. The earlier study focused on utterance-final monosyllabic accent groups, produced with a single "high" pitch accent, a low phrase accent, and a low boundary tone (Pierrehumbert label H*LL% [5]; Figure 1, left panel). The current data base also includes H*LL% contours for polysyllabic accent groups, continuation contours (H*LH%), and Yes/No contours (L*H%). Continuation contours consist of a dual motion in which an early peak is followed by a valley and a final rise (Figure 1, center panel). Yes/No contours (Figure 1, right panel) consist of a declining curve for the pre-accented region (not shown), an accelerated decrease starting at the onset of the accented syllable, and then a steep increase in the nucleus. Unless stated otherwise, results are reported for H*LL% contours.

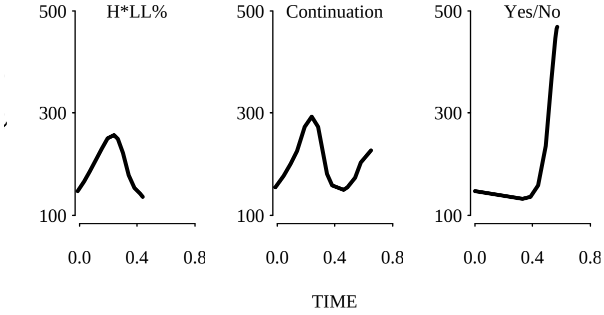

Figure 1: **Averages of H*LL%, Continuation, and Yes/No contours.**

2.1. Effects of accent group duration

As point of departure we take the most obvious analysis: *Measure alignment of H*LL% accent curves in terms of peak location,* and *assume that accent group structure can be captured simply by total duration.* There is indeed a statistically significant correlation between peak location and total duration, showing that peaks are not placed either a fixed or random millisecond amount into the stressed syllable. But the correlation is weak (0.57). We could stop here, and declare that accent curve timing is only loosely coupled to accent group structure. Or, as we do next, we can measure whether timing depends on aspects of accent group structure other than total duration.

2.2. Effects of segmental structure

In [6] it was shown that peak location strongly depends on segmental structure. For monosyllabic accent groups, peak location (measured from accented syllable start) is systematically later in sonorant-final accent groups than in obstruent-final accent groups (pin vs. pit), and later in obstruent-initial accent groups than in sonorant-initial accent groups (bet vs. yet). Such effects persisted when we measured peak location from vowel start instead of syllable start, and when we normalized peak location by division by syllable or rhyme duration. Apparently, peaks are located at neither a fixed millisecond amount nor a fixed fraction of the accent group.

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

2.3. Effects of accent group “sub-durations”

While these data undermine most peak placement rules used in text-to-speech synthesis, they do not unambiguously disqualify the overall accent group duration hypothesis: overall duration tends to be longer for “pin” than for “pit”, and longer for “bet” than for “yet”; in addition, the hypothesis does not require that peaks are located at a fixed fraction into the accented syllable or its rhyme. A better test concerns the prediction that changes in peak location do not depend on which “part” of an accent group is lengthened. To illustrate, when we contrast two renditions of the same two-syllable accent group that have the same overall duration of 400 ms, but the durations of the syllables change from 210+190 ms to 250+150 ms, is peak location the same? Or does the 40 ms lengthening of the first syllable have a larger effect than the 40 percent shortening of the second syllable?

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

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

Here, a is a rendition of an accent group with segmental structure **S**, $T_{peak}(a)$ is peak location, j refers to the j-th “part” of the accent group, $D_j(a)$ is the corresponding duration, and $\alpha_{\mathbf{S},j}$ its weight. Lacking space for detailed definitions, we use three “parts”: accented syllable onset, accented syllable rhyme, and remaining unstressed syllables (polysyllabic accent groups only).

For all three contour classes, the weights $\alpha_{\mathbf{S},j}$ vary strongly as a function of part location (j = *onset, rhyme, remainder*), with the effects of the onset being the strongest and the effects of the remainder being the weakest. Thus, the peak is later in the 250+150 ms rendition than in the 210+190 ms rendition, thereby conclusively disproving the overall accent group duration hypothesis. Effects of segmental structure (only analyzed for the H*LL% class) were weaker, and virtually absent for onsets, indicating that the effects of onset (“bet” vs. “yet”) on peak location are largely due to intrinsic duration differences between onsets (e.g., /b/ is longer than /y/). Setting the intercept $\mu_{\mathbf{S}}$ to zero did not affect the fit, indicating that the *accented syllable start plays a pivotal role in alignment,* and not vowel start.

2.4. Anchor points

Estimation of Anchor Points The peak is only one point on an accent curve, and it is not clear whether it is the perceptually most important point – perhaps the rise is. One way to get a handle on the entire curve is by measuring and predicting selected points on that curve (“anchor points”). Towards this end, we subtract a locally straight “phrase curve” from the observed F_0 curve around the area where the accent curve is located, and then consider the residual curve as an estimate of the accent curve. We then sample the estimated accent curve at locations corresponding to 5%, 10%, 25%, etc. of maximal height. The corresponding points are the anchor points.

We estimated the locally straight “phrase curve”, with an optimization algorithm that minimized the weighted least squares deviation (in the logarithmic domain) between the observed F_0 contour and the sum of a straight local phrase curve, a warped accent curve template, and perturbation curves (see below).

Obviously, the model in Equation (1) can be applied to any anchor point by replacing the *peak* subscript by i, for the i-th anchor point:

$$T_i(a) = \sum_j \alpha_{\mathbf{S},j} \times D_j(a) + \mu_{\mathbf{S}}. \qquad (2)$$

We call the ensemble of regression weights the *alignment parameter matrix*, and this equation the *alignment model*.

Alignment model and time warping The alignment model is equivalent to the statement that individual accent curves for a given pitch accent class *are obtained from a common template via a parameterized time warp*. The template consists of n pairs $< i,\ P_i >, i = 1, \cdots, n$; i is the index of the anchor point at P_i percent of maximal height. The time warp for accent group rendition a is $Warp_a(i) = T_i(a)$, and maps template time (i) onto the time axis in recorded speech. Thus, the family of all warp functions associated with an accent curve class is defined to consist of those functions that are produced by the alignment parameter matrix for that class.

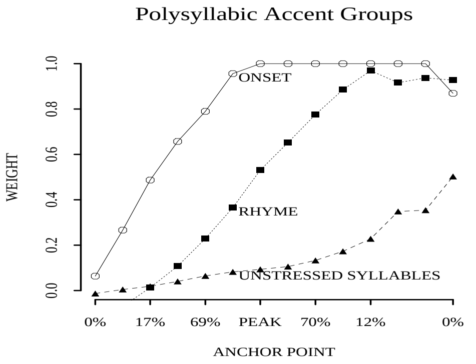

Figure 2: **Regression weights as a function of anchor point, for each of the three sub-intervals of the accent group. H*LL% accent type. Solid curve: onset; dotted curve: rhyme, dashed curve: remainder.**

Alignment parameter results Figure 2 shows the values of the alignment parameters for polysyllabic phrase-final accent groups (H*LL%). We note the following. First, the overall fit is quite good (predicted-observed correlation of 0.91 for peak location), explaining more than 2.5 times the variance explained by overall accent group duration (correlation of 0.59) Second, the weights for the onset exceed the weights for the rhyme, and the latter exceed the weights for the remainder of the accent group. In other words, lengthening the onset duration of the stressed syllable has a larger effect on any location of the accent curve than lengthening the duration of the unstressed syllables. Third, the curves are monotonically increasing. They initially diverge, and then converge. Early anchor points mostly depend on onset duration and hardly on the durations of the rhyme and the remainder, but late anchor points depend more evenly on all three subsequence durations. A key point is that these alignment curves are well-behaved, and without a doubt can be captured by a few meta-parameters (e.g., two straight line segments per curve).

3. Proposed pitch Model

For generating F_0 curves we have to specify where an accent curve is to be placed on the vertical (frequency) axis, how its height is controlled, how it is tilted, and how to combine successive accent curves. It is not clear how to do this in the tonal interpolation (or "linear") tradition (e.g., [5]), but quite clear in the superpositional framework (e.g., [4]. In the superpositional tradition, vertical placement is accomplished by adding accent curves to a phrase curve. Combination of successive accent curves follows as a side effect of this addition. We think, however, that unless the claims of these two traditions are made more specific and mathematically precise, it is very difficult to empirically distinguish between them. It is with these relativistic remarks in mind that we propose to use a superpositional approach, which we now outline.

3.1. Additive decomposition

In the best-known superpositional model, the Fujisaki model [1, 4], the observed F_0 curve is obtained by adding (in the logarithmic domain) three curve types with different temporal scopes: Phrase curves, accent curves, and a horizontal line representing the speaker's lowest pitch level. We likewise propose to add curves with different temporal scopes, but remove the base pitch line and include segmental perturbation curves instead (see below).

Phrase curves We found that phrase curves could be modeled as two-part curves obtained by (non-linear) interpolation between three points: the start of the phrase, the start of the last accent group in the phrase, the end of the phrase. This alignment is analogous to that of accent curves: accent (phrase) curves are aligned in terms of selected accent-group-internal (phrase-internal) segment (accent group) boundaries. The phrase curve model includes as special cases the standard (linear) declination line, and curves that are quite close to the phrase curve in Fujisaki's model. We prefer to be open to the possibility that phrase curves exhibit considerable and meaningful variability. Just to make a conceptual point, one can account for a plateau-like curve ("hat pattern") bounded by accent lending rise and fall by making the phrase curve bulge upward and then downward somewhat more strongly than in the Fujisaki model, and positioning two accent curves at the rise and fall locations; the second accent curve is negative.

Curve parameters are controlled by sentence mode and locational factors (e.g., sentence location in the paragraph). We are also considering hierarchical possibilities [3].

Perturbation curves Perturbation curves are associated with initial parts of sonorants following a transition from an obstruent. We measured these effects, by contrasting vowels preceded by sonorants, voiced obstruents, and unvoiced obstruents in deaccented syllables [6]. These curves are described by a rapid decay from values of about log(1.4) to 0 in 100 ms, and are added in the logarithmic domain to the other curves.

3.2. Accent curve height

In our model, accent curve height is determined via a multiplicative model by multiple factors, including position (in the minor phrase, the minor phrase in major phrase, etc.) and factors predictive of prominence. Thus, height can have many values, and is not itself "phonological".

3.3. General assumptions of the model

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

$$F_0(t) \;=\; \bigoplus_{c \in C} \bigoplus_{k \in c} f_{c,k}(t). \qquad (3)$$

C is the set of curve classes (e.g., $\{$*perturbation, phrase, accent*$\}$), c is a particular curve class (e.g., *accent*), and k is an individual curve (e.g., accent curve). The operator $\bigoplus$ 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 ($a \oplus b = b \oplus a$). A key assumption is that each class of curves, c, corresponds to a *phonological entity with a distinct time course*. For example, the *Phrase* class has a longer scope than the *Accent* class.

A central issue to be resolved for models in this class is which parameters of which curve classes depend on which factors. For example, in our model the alignment parameters do not depend on any phrase-level factors, and the perturbation curves are completely invariant.

Sub-duration directional invariance. In the same way as addition of curves in the log domain is only a special case of a much more general decomposition principle (Eq. 3), the linear alignment model is a special case of what we call the *sub-duration directional invariance* principle, according to which for any two accent groups a and b that contain segmentally equivalent parts:

$$If\ D_j(a) \geq D_j(b)\ for\ all\ j\ then\ T_i(a) \geq T_i(b). \qquad (4)$$

Our alignment model is a special case, because when $D_j(a) \;\geq\; D_j(b)$ for all j, then $\sum_j \alpha_{\mathbf{S},j} D_j(a) \;\geq\; \sum_j \alpha_{\mathbf{S},j} D_j(b)$, and hence $T_i(a) \geq T_i(b)$.

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

Generalized accent groups We could generalize the concept of accent group, which is based on syllables being dichotomized into stressed and unstressed syllables. For example, we could trichotomize syllables into Strong, Medium, and Weak, and posit that there are two types of accent groups, Strong and Medium, that might overlap (share syllables). Strong accent groups would start with strong syllables and be terminated by strong, but not by medium or weak, syllables; medium accent groups would start with medium syllables and be terminated by either strong or medium, but not by weak, syllables.

4. Conclusions

This paper presented data on alignment that must be accounted for by any intonation model claiming to describe both the fine and coarse details of observed F_0 curves. The proposed alignment model provides a very good fit, but we reported no analyses excluding alternative accounts.

To return to the basic question asked in the Introduction, the model captures variation of accent curves belonging to the same phonological class in terms of a shared template and matrix of alignment parameters. This linear model is obviously only a first-order approximation; different models with the sub-duration directional invariance property need to be considered. In addition to exploring different models, a wealth of perception studies are suggested by the model.

While the Fujisaki model plays a central role in the linear vs. superposition controversy, it is clearly not the only model that one can call superpositional. We believe that it is important to focus on broader properties of the superposition concept; the framework sketched in the preceding section may serve as a first step to make these broader properties clearer.

5. REFERENCES

1. H. Fujisaki. Dynamic characteristics of voice fundamental frequency in speech and singing. In Peter F. MacNeilage, editor, *The production of speech*, pages 39–55. Springer, New York, 1983.
2. K.J. Kohler. Macro and micro f0 in the synthesis of intonation. In Kingston J. and M. E. Beckman M.E., editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pages 115–138. Cambridge: Cambridge University Press, 1990.
3. D. Robert Ladd. Declination 'reset' and the hierarchical organization of utterances. *Journal of the Acoustical Society of America*, 84:530–544, 1988.
4. B. Möbius, M. Pätzold, and W. Hess. Analysis and synthesis of german f0 contours by means of fujisaki's model. *Speech Communication*, 13, 1993.
5. Janet Pierrehumbert. *The phonology and phonetics of English intonation*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1980.
6. J. P. H. van Santen and J. Hirschberg. Segmental effects on timing and height of pitch contours. In *Proceedings ICSLP '94*, pages 719–722, 1994.