3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SEGMENTAL EFFECTS ON TIMING AND HEIGHT OF PITCH CONTOURS

Jan P. H. van Santen and Julia Hirschberg

Linguistics Research Department, AT&T Bell Laboratories
Murray Hill, NJ, 07974, U.S.A.

ABSTRACT

Instantiations of a given type of pitch contour vary systematically as a function of the segmental material with which the contour is associated. In this study, we report on results of statistical analyses of 1,926 utterances collected from a single female speaker. Results show that the observed pitch contours can be quantitatively modeled as distortions in the temporal and frequency domains of a single, underlying contour. Distortions in the temporal domain can be described as non-linear rightward stretching of the contour as the durations of onset, vowel nucleus, and coda increase. Variations in the frequency domain include effects of vowel height and overall syllable duration. Frequency effects of onset voicing can be attributed to systematic differences in duration between voiced and voiceless consonants as well as to a small local perturbation confined to the first 50 ms after the vowel onset.

I. INTRODUCTION

In natural speech, F_0 contours for nominally the same category of pitch contour vary as a function of the segmental material with which the pitch contour is associated. Many of these effects have been studied (e.g. [3]), but usually only one or two factors (e.g.,vowel height) or only one aspect of the contour (e.g., peak location) were addressed. What is lacking is a quantitative description of the *joint effects* of segmental factors on the time course of an *entire* pitch contour. This paper attempts to provide such a description.

The theoretical importance of modeling segmental effects is that it may cast light on the general issue of the relation between articulation and pitch control. While we will not directly address this here, we hope that a good descriptive model will contribute to its resolution.

The practical importance concerns text-to-speech synthesis. It has often been assumed that segmental effects on pitch are too small to be perceived (e.g., [4]). However, we believe that our understanding of perception of pitch in fluent, meaningful speech is currently not sufficient to make strong claims about the *im*perceptibility of any aspect of speech, so that currently we have no other option but to model any effect on any acoustic feature that can be clearly demonstrated in natural speech. Moreover, there is a rising body of evidence suggesting that some segmental effects are not only perceptible but that modeling them adds to both naturalness and intelligibility of synthetic speech (e.g., [3]).

II. METHOD

Data collection

A female speaker was used for all productions. Recorded, digitized speech was segmented manually and labeled for pitch accent, phrase accent and boundary tone type, using the system of intonational description presented in Pierrehumbert [2]. An epoch based pitch tracker [5] measured pitch at 10 ms intervals.

Our focus is on simple phrases where a phrase final *target syllable* receives nuclear stress. The phrases were produced as single intonational phrases, with a single H* pitch accent, a low phrase accent, and low boundary tone, and thus would be described as H*LL% in Pierrehumbert's system. We also analyzed sentences containing a deaccented target syllable.

For the first case (1727 tokens), sentences were of the type: "Now I know *X*", where *X* is the accented target syllable. In the second case, three phrasal locations for the target were used: "Now I know X I *mean*" (68 tokens), "I *mean* now I know X *more*" (69 tokens), and "I *mean* now I know X" (62 tokens).

Target syllables had the form $C_o v C_c$, where the onset C_o and the coda C_c consisted of zero or more consonants and where v was a single vowel or diphthong.

Definitions

Conventions. Results reported on effects of consonants will be stated in terms of broad *consonant classes*: *- V* (voiceless consonants), *+ V - S* (obstruent, voiced consonants), and *+ S* (sonorants). In these recordings, vowels in vowel-initial syllables were invariably preceded by a glottal stop; this stop is assigned to class + V - S. Empty codas (i.e., open syllables) are assigned to class + S. Vowel classes are *high*, *mid*, and *low*.

The class of an onset or coda is defined as that of its least sonorous consonant; thus, the syllable "blank" has *onset class* + V - S and *coda class* - V.

As we shall see, an important role is played by the end of the last sonorant in the target syllable (the end of /n/ in "blank", the end of /i/ in "seat" and "see"). We refer to this point as the *sonorant end*. The corresponding time measured from syllable start is denoted T_{s-end}.

The *sonorant rhyme*, or *s-rhyme*, is defined as the interval that starts at the start of the last sonorant in the onset (or vowel start if the onset has no sonorants) and that ends at the sonorant end. Thus, the s-rhyme is "lan" in "blank", /i/ in "seat", /yu/ in "muse", and /in/ in "seen". The duration of the s-rhyme is denoted $D_{s-rhyme}$.

The onset duration, D_{onset}, is the duration of the interval from syllable start to the start of the s-rhyme (/b/ in "blank", /s/ in "seat", and /m/ in "muse"). Hence, $T_{s-end} = D_{onset} + D_{s-rhyme}$.

Underlying contour, phrase curve, anchor point. Figure 1 shows contours for the three different onset classes, and illustrates the concept of "underlying" contour. All-sonorant syllables are often assumed to provide an unobscured picture of the "true" underlying contour – the contour as intended by the speaker. Syllables with - S onsets leave out the section corresponding to the - S onset and, in addition, contain a perturbation of the initial part of the s-rhyme. In Figure 1, we filled in these missing or perturbed sections (dashed curves).

We found that a straight line (dotted lines in Figure 1) drawn between the point 70 ms prior to syllable onset (in the center of the [O] in "know") and the sonorant end and subsequently subtracted from the contour produces a contour that is statistically more stable than the original. Following the common idea of decomposing a pitch contour into a phrase curve and an accent curve (e.g., [1]), we consider this straight line as being part of a "phrase

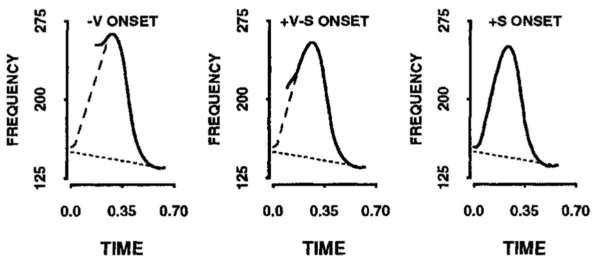

Figure 1: **Averaged contours for syllables with - V, + V - S, and + S onsets and + S codas. Dotted line: local phrase curve; dashed line: estimated "underlying" contour.**

curve" that spans the entire utterance. Note that we make the minimalist assumption that the phrase curve is *locally* (i.e., within the target syllable), but not globally, linear.

This straight line and the pitch contour in the remainder (carrier part) of the phrase vary between utterances, but this variation is strikingly independent of the target syllable. The pitch value that is reached 70 ms prior to target syllable onset is constant (it ranges from 149 to 151 Hz across onset classes and from 149 to 150 across coda classes.) The *sonorant* end point has a 5-Hz range (137, 134, and 132 Hz for - V, + V - S, and + S codas). Moreover, this already small range has been inflated by the fact that (1) the average slope of the line between the 70 ms pre-syllable point and the sonorant end point is negative, and (2) the sonorant end is much later for + S codas (644 ms) than for + V - S (460 ms) or - S codas (362 ms). By contrast, the value reached at the *vowel* end is far more variable (169 Hz for + S codas, 134 and 137 Hz for + V - S and - V codas). We interpret this as the speaker returning to the phrase curve at the end of the last sonorant in the syllable, not at the end of the vowel.

The assumption of a phrase curve leads to a very useful concept, that of *anchor point*. The *20 percent pre-peak anchor point* is defined as the point that precedes the peak and whose value (minus the value of the phrase curve) is 20% of the peak value (minus the value of the phrase curve). The peak itself can be described as the 100 percent anchor point. We found that these anchor points are much better behaved than anchor points defined otherwise.

III. EFFECTS ON TIMING

Effects of onset and coda class

Most results in this section concern peak timing, but results for other anchor points will also be briefly discussed. Figure 2 shows peak timing as a function of onset and coda class, restricting the analysis to mid vowels. The main point is that peak location is not invariant across onset or coda classes, whether it is measured from syllable onset or vowel onset, or is normalized by division by sonorant end time or s-rhyme duration. The consistency of the effects of coda class across onset classes and vice versa shows that the effects are highly systematic.

One of the obvious differences between consonant classes are durational: - V consonants are longer (173 ms) than + V - S (125 ms) or + S (104 ms) consonants, and, as noted above, the sonorant end is much later for + S than for + V - S or - V codas. Thus, the effects seen in Figure 2 could be summarized by stating that peak time increases with D_{onset} and $D_{s-rhyme}$. However, the effects are not simple because there is not a rigid shift (the curves in the top right panel are not horizontal) nor a proportional stretch (the curves in the bottom two panels do not coincide). In addition, it seems a priori unlikely that effects of onset and coda are purely due to their durational differences, because there are other and far more profound differences between consonant classes.

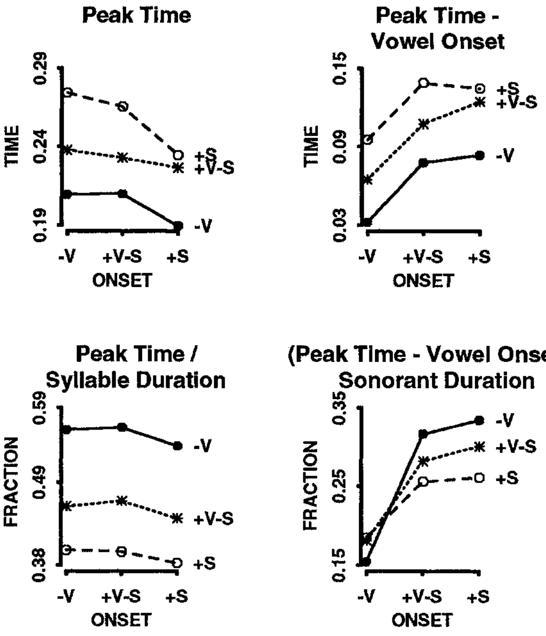

Figure 2: **Effects of onset and coda (curve parameter) on peak time.**

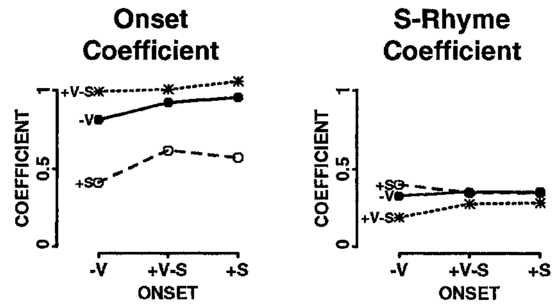

Figure 3: **Regression coefficients for onset duration (α; left panel) and s-rhyme duration (β; right panel) as a function of onset and coda (curve parameter).**

Effects of within-consonant durational variation

A more instructive way to analyze these temporal effects involves the following linear model:

$$T_{peak}(D_{onset}, D_{s-rhyme}; C_o, v, C_c) = \alpha_{C_o,C_c} \times D_{onset} + \beta_{C_o,C_c} \times D_{s-rhyme} + \mu_{C_o,C_c}. \quad (1)$$

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.

In terms of this model, the data in Figure 2 have multiple interpretations. On the one hand it could be that all weights are zero so that the data directly, and only, reflect the values of the μ parameters. On the other hand the latter may be zero, the weights may be constant (independent of C_o and C_c), and the data in the Figure may be the exclusive result of duration differences between consonant classes.

The model can also be used to analyze the validity of some very simple rules about the invariance of peak location:

1. Measured from syllable onset: $\alpha = \beta = 0$.

2. Measured from vowel onset: $\alpha = 1$, $\beta = 0$.
3. Measured from syllable onset, divided by T_{s-end}: $\alpha = \beta$.
4. Measured from vowel onset, divided by $D_{s-rhyme}$: $\alpha = 1$.

The intercept, μ, proved to be a statistically insignificant parameter whose presence added less than 2 percent to the amount of variance explained. The results (Figure 3) were based on the model altered by removing the μ parameters. Correlations between observed and predicted peak locations varied between 0.61 and 0.87, with a median of 0.77. The median absolute deviation between predicted and observed peak location was 15 ms.

Figure 3 shows that the α parameters were invariably larger than the β parameters, and that the latter were well above zero (all statistical significance levels are at $p < 0.001$ and will not be reported individually). These observations contradict the first three rules. The fourth rule fails completely for + S codas.

We now turn to the issue of whether the effects shown in Figure 2 are due exclusively to the durational differences between consonant classes. The parameters vary as a function of coda, and much less (if at all) as a function of onset. Somewhat to our surprise, this means that the effects of onset class shown in Figure 2 can be explained purely in terms of durational differences. But the effects of coda do involve differences other than durational ones.

The same model was applied to anchor points other than the peak (positioned at 5, 20, 50, 80, and 90 percent of the peak height [after subtraction of the phrase curve] on both sides of the peak), producing similar results. In addition, we found that weights increase for later anchor points.

IV. EFFECTS ON FREQUENCY

Consonantal perturbation

As stated in Section II, we found no effects of consonants on pitch values of the preceding vowel. By contrast, there are strong effects on the following vowel. Figure 4 shows pitch during the first 150 ms after vowel onset. Because of somewhat limited quantities of data for the deaccented case, the values for both cases were corrected by fitting the following model:

$$Freq_{obs}(t;\ C_o, v, C_c) = Freq_{true}(t; C_o) + Freq_{true}(t; v) + Freq_{true}(t; C_c) \quad (2)$$

The parameters in the right hand of the equation can be estimated with standard least squares methods [6]. For the accented case, no differences were observed between "raw" and corrected pitch values. The values of $Freq_{true}(t;\ C_o)$ represent the time course for onset class C_o corrected for the effects of vowel height and coda class.

There is a clear consonantal perturbation effect for both accented and deaccented syllables. It is short-lived, and all but vanishes after 50 ms.

We also performed the analysis separately for deaccented syllables in the three phrasal positions ("Now I know X I *mean*", "I *mean* now I know X *more*", and "I *mean* now I know X"), and found similar results (which allowed us to pool the results). This is of interest because the duration of phrase-final syllables was more than 70 percent longer than that of syllables in the other phrasal locations. Thus, the time course of consonantal perturbation is not affected by factors that have strong effects on peak location. This means that any attempt to model the accent component must have at least two sub-components, each with its own time course being affected by different segmental factors.

The consonantal perturbation effects appear to be larger for accented syllables. However, as Figure 1 showed, vowels following - S onsets start at a later point where the hypothetical underlying curve has already reached higher values. Thus, the results in the right panel reflect both consonantal perturbation and duration differences between onset classes.

Vowel height

It is well-known that high vowels tend to produce higher pitch values than low vowels. Figure 5 shows pitch contours for high, mid, and low vowels, during the first 150 ms after vowel onset. As in the case of consonantal perturbation effects, we used corrected frequencies $[Freq_{true}(t; v)]$. Results indicate that vowel height had no effect for deaccented syllables, but had at least a 15 Hz effect for accented syllables. Yet, there was no evidence for vowel reduction, possibly due to the slow speaking rate.

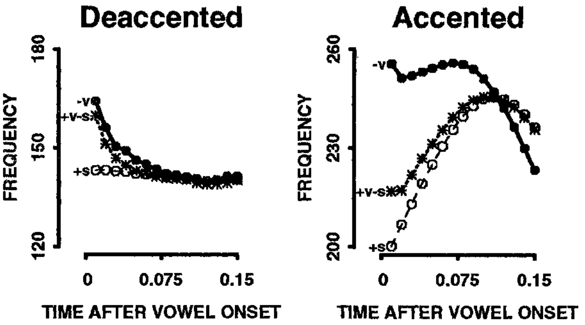

Figure 4: **Effects of onset class on pitch during the first 150 ms after vowel onset for deaccented (left panel) and accented (left panel) syllables.**

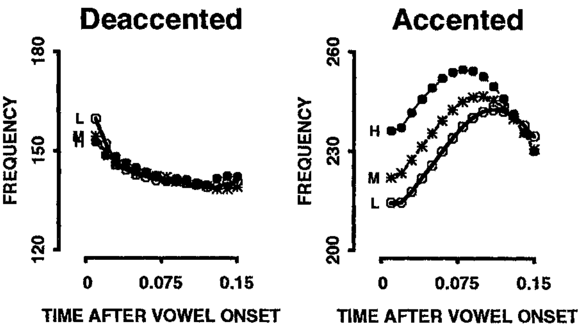

Figure 5: **Effects of vowel height (L=low, M=mid, H=high) during the first 150 ms after vowel onset for deaccented (left panel) and accented (left panel) syllables.**

One way to capture the difference between accented and deaccented syllables is by positing that vowel height acts as a multiplyer on the difference between the contour and the phrase curve. Since, for deaccented syllables, the contour coincides with the phrase curve, there is no effect of vowel height.

Overall duration

It is to be expected that the excursion as measured, e.g., by the difference between the peak height and the local phrase curve, is correlated with overall syllable duration or sonorant end time. One reason is that pitch velocity (changes in pitch value) has an upper bound, so that very short durations cannot involve large pitch excursions. We posit that if any relation exists it can be described by a *sigmoid* curve. This means that for very short syllables the accent curve coincides with the phrase curve until some critical duration is reached; and that there is some upper limit beyond which further increases in duration have little effect on excursion.

Indeed, correlations between peak height and sonorant end time were highly significant (0.43, 0.43, and 0.38 for - V, + V - S, and + S codas, respectively). We fitted sigmoid curves for all nine combinations of onset and coda classes, given by:

$$Peak(T_{s-rhyme}; C_o, C_c) = \\ LO(C_o, C_c) + \frac{[HI(C_o, C_c) - LO(C_o, C_c)]}{[1 + e^{-slope_{C_o,C_c}(T_{s-rhyme} - loc_{C_o,C_c})}]}. \quad (3)$$

Here, $LO(C_o, C_c)$ and $HI(C_o, C_c)$ are the low and high asymptotes of the curve, *slope* is the slope at the point of steepest ascent, and *loc* the location of this point. The only parameter that differed significantly among the nine combinations was the location parameter, and this parameter depended only on the coda class (with values of 0.26, 0.43, and 0.58 for - V, + V - S, and + S codas, respectively).

V. MODEL OF SEGMENTAL EFFECTS ON PITCH

Formal description

We now present a model that combines the findings reported thus far. The model extends the weighted linear model in Eq. (1). The effects of vowel height and overall duration on peak height are modeled as multiplicative effects on the difference between contour and phrase curve, *DFreq*, while the effects of consonant perturbation are modeled as an additive effect. Note that in the temporal domain, the latter depends only on time after vowel onset. In it most general form, the model is given by:

$$DFreq[t; D_{onset}, D_{s-rhyme}, C_o, v, C_c)] = \\ Perturb[t - D_{onset})] \times A[C_o, v, C_c] + \\ \frac{Hght[v] \times ULC[Warp(t; D_{onset}, D_{s-rhyme}, C_o, v, C_c)]}{[1 + e^{-slope_{C_o,C_c}(T_{s-rhyme} - loc_{C_o,C_c})}]}. \quad (4)$$

Here, $DFreq()$ is the contour after subtraction of the phrase curve. *Perturb* describes the perturbation generated by the onset consonant and corresponds to subtracting the + S curve from the - V curve in Figure 4, left panel. A is a parameter modulating the magnitude of the perturbation. Our results reported above suggest that A only depends on C_o and that $A[-S, v, C_c] = 0$. $Hght[v]$ is the multiplicative effect of vowel height. The effect of overall duration involves the the sigmoid curve specified in Eq. (3). Preceding results suggest that *slope* is constant, and that *loc* depends only on C_c. ULC is the underlying curve, specified by 14 sampled time values (ranging between time=0 and time=1) and corresponding frequency values (likewise normalized to range between 0 and 1.)

Finally, for any token, $Warp(t; D_{onset}, D_{s-rhyme}, C_o, v, C_c)$ defines a *time warp function* that maps t values between 0 and T_{s-end} to values between 0 and 1 (i.e., the time interval of ULC). This mapping is computed as follows. Consider the i-th point of ULC, corresponding to some time point t_i in the interval (0,1). This point corresponds to the point $t(t_i)$ in the interval $(0, T_{s-end})$ given by:

$$t(t_i) = \alpha_{i,C_o,C_c} \times D_{onset} + \beta_{i,C_o,C_c} \times D_{s-rhyme} \quad (5)$$

In other words, $t(t_i)$ is the location in the token that corresponds to the i-th anchor point. Thus, when i is the peak point of ULC, then $t(t_i)$ is simply the *predicted peak location*. Once the mapping is computed for all anchor points i, the mapping is completed by linear intrapolation.

Results

The model was fitted with a variety of *dependency patterns*, i.e., assumptions about which parameters depend on both, one, or neither of C_o and C_c. The results mirrored the effects observed in the preceding sections. Thus, we found little evidence for an effect of C_o on the α and β parameters; *slope* was constant, and *loc* depended only on C_c; A only depended on C_o.

The overall fit of the final model (i.e., the model with the dependency pattern just described) was 15.7 Hz (root mean squared deviation, or *rms*; median absolute deviation: 10.2 Hz). Here, the comparison was over all frames in sonorant regions of the target syllable. Confining the comparison to only the peaks, the rms deviation was 17.7 Hz (median absolute deviation: 12 Hz). The correlation was 0.508. In the temporal domain, corresponding values were 20 and 10ms, and a correlation of 0.87.

To establish a baseline, we formed groups of tokens with identical onset class, vowel height class, coda class, and (within 10 ms) segmental durations. In other words, all tokens in such a group would get identical predicted contours. The variability within groups is a measure of the limits of any attempt to predict these contours. Averaged over these groups, the standard deviation of peak height and location was 9.75 Hz and 14 ms, respectively. Thus, the values of 17.7 Hz and 20 ms are only 80 (and 40) percent larger than what can be considered the theoretical minimum.

VI. CONCLUSION

We analyzed segmental effects on pitch contours. The key effects in the frequency domain were those of vowel height, consonant-induced perturbations, and overall syllable duration. In the temporal domain, we showed that timing of "anchor points", such as the peak, can be predicted from segment duration in conjunction with coda class; no information is needed about onset class or vowel height. These temporal effects cannot be further reduced to simpler rules, such as the rule that peak location is some fixed percentage of overall syllable duration.

More abstractly, we have decomposed the observed pitch contours into three components that each have their own time course and are influenced by different factors. The consonant-induced perturbation is the most local and only depends on onset class and time elapsed since vowel onset. The phrase curve is assumed to be largely independent of segmental effects and of accent type. The accent contour is described as a deviation from the phrase curve; this deviation is modulated by vowel height and by overall syllable duration. In the temporal domain, the contour is warped non-linearly, with timing of early parts depending largely on onset duration while timing of later parts are also increasingly more influenced by the s-rhyme duration.

The predictability of the timing of anchor points suggests a rather close coordination of segmental and pitch control. Of course, it remains to be seen if similar results can be obtained in more varied text materials.

Extending the model to other classes of pitch contours should be relatively straightforward; we have already done so for yes/no question contours and for H*LL% contours with nuclear stress on polysyllabic words. What will be more challenging are contours with multiple pitch accents, especially when these are in close proximity to each other.

References

[1] H. Fujisaki. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Fujimura, editor, *Vocal physiology: voice production, mechanisms and functions.* Raven, New York, 1988.

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

[3] K. Silverman. *The Structure and Processing of Fundamental Frequency Contours.* PhD thesis, Cambridge University, Cambridge UK, 1987.

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

[5] D. Talkin and J. Rowley. Pitch-synchronous analysis and synthesis for tts systems. In *Workshop on speech synthesis,* pages 55–59, Autrans France, 1990. ESCA.

[6] J. P. H. van Santen. Contextual effects on vowel duration. *Speech Communication,* 11:513–546, 1992.