Prosodic Peak Estimation under Segmental Perturbations.

Greg Kochanski

Oxford University Phonetics Laboratory,

Oxford,

UK

(Dated: October 19, 2009)

Abstract

Despite the apparent simplicity, measuring the position of peaks in speech fundamental frequency (f_0) can produce unexpected results in a model where f_0 is the superposition of an supersegmental component and a segmental component. In these models, the measured f_0 peak position can be as much as an entire syllable different from the peak of the intonation component. This difference can be large enough so that the measured peak positions could falsely suggest a phonological distinction in the intonation where none really exists. This paper then discusses measurement techniques that are less sensitive to segmental effects than directly measuring the position of the f_0 maximum. A algorithm, called the "bracketed maximum" is presented. The performance of these techniques is compared on a corpus of speech data where the intonation is expected to be in a stable position. The bracketed maximum can reduce the variance of peak position measurements by at least 15%, in the presence of changing segmental structure, thereby presumably yielding a more accurate measurement of the intonation peak position.

PACS numbers: 43.60.-c, 43.70.Jt, 43.72.Ar

Phonetics Laboratory, 41 Wellington Square, Oxford OX1 2JF, UK

I. INTRODUCTION

Many papers on intonation are based upon measurements of the timing of peaks in fundamental frequency (f_0) contours (e.g. Chen *et al.* (2004), House (2003), Ladd *et al.* (1999), Arvaniti *et al.* (1998), Silverman and Pierrehumbert (1990), and Pierrehumbert and Steele (1989)). Peak timing is an easy, objective measurement, but f_0 is generally considered to be jointly determined by the segments of the utterance ("microprosodic perturbations") and suprasegmental prosodic properties. This paper investigates obtaining timing of underlying prosodic peaks from measurements on f_0 curves.

This paper has two objectives: first, to show that separating segmental and suprasegmental effects is not straightforward, even in a simple "toy" model: there are unexpectedly large nonlinear interactions. Second, to compare several different ways of measuring the peak timing to find which is best at estimating the timing of the prosodic peaks.[1] We look for the measurement technique that is least affected by changes in the segmental structure of a sentence.

This paper assumes a superposition model for intonation. Superposition models assume that an observable is made of a sum of two components and that the the two components are independent of each other. One can always split an observed time-series into two or more components; in fact, there are an infinite number of ways to do so. However, a superposition model requires more than just the addition of two components; it also asserts that the two components are independent and that they can be combined in any desired way.[2] This provides some constraint on how f_0 can be split into two components.

Another constraint comes about because intentional control of f_0 is not fast enough to reproduce or compensate for segmental effects. The muscles of the larynx cannot produce an increase of f_0 in less than 100 ms (Stevens, 1998 pp. 40–48 and references therein, Xu and Sun, 2000), while segmental effects can come and go in that interval. This means that,

to some level of approximation, we can split the segmental effects away from the prosodic part by their characteristic time scale: short-term effects can be treated as aerodynamic or segmental effects and changes over time-scales longer than 100-200 ms are plausibly prosodic. Such a division is consistent with the linguistic association of prosody with suprasegmental properties (i.e. properties spanning more than one segment).

For f_0, the first component in a superposition model is the prosodic contour, which is normally assumed to be determined by the choice of accents and their positions. The second component is the segmental effect: a change in f_0 that depends on the phone. Low vowels characteristically have a low f_0, whereas high vowels have a somewhat higher f_0 on average (Crandall, 1925; Taylor, 1933; Peterson and Barney, 1952; House and Fairbanks, 1953; Ladefoged, 1964; Whalen and Levitt, 1995). This vowel-dependent shift in f_0 is variously estimated to be from 4 Hz to above 10 Hz, i.e. roughly 0.5 semitones. While this shift is fairly small compared to the (roughly) 3 semitone standard deviation of f_0 in normal speech (Baken, 1987, Table 5-2), it will be seen that it is not small enough to safely ignore.

Consonants can also affect f_0. For instance, nasality has been observed to have a significant effect (Silverman, 1987). Also, van Santen and Hirschberg (1994) showed that voiceless consonants and voiced obstruents can make short-lived changes of about 20 Hz near the onset of voicing. Thus the size of these consonant effects is comparable to those caused by vowel height.

One example of a superposition intonation model is Fujisaki (1983), who constructed f_0 contours from a superposition of short-term and long-term prosodic effects. Other models of intonation that add segmental effects onto a prosodic component are Morlec *et al.* (1996), Di Cristo and Hirst (1986), van Santen and Hirschberg (1994)[3], van Santen and Möbius (1999) and Ross and Ostendorf (1999).

A superposition model for intonation can be written as

$$f(t) = p(t) + s(t), \tag{1}$$

where $f(t)$ is the observed (surface) frequency at time t, $s(t)$ is the segment-related frequency

shift, and $p(t)$ is the prosodic (accent-related) part. Superposition implies that for any $p(t)$ which is a possible prosodic contour, and for any $s(t)$ which is a possible segmental frequency shift, then $p(t) + s(t)$ is a valid, physically possible intonation contour.[4]

The independence assumption in superposition models is probably just an approximation. But, if it were not a useful approximation, one would expect to see a substantially different choice of segments in regions of low f_0 vs. high f_0, or alternatively, substantially different phonetic implementation of segments in low f_0 regions from high f_0 regions. While there are differences in segmental content of accented vs. non-accented regions (e.g. schwa is rare under accented regions or see Greenberg *et al.*, 2001), the correlation with f_0 is presumably not strong, because high f_0 is not strongly correlated with accent/prominence (Kochanski *et al.*, 2005; Kochanski and Orphanidou, 2008).

Also, the fact that segmental effects have been measured suggests that they are not compensated by speakers: if speakers automatically compensated for $s(t)$, then one would not expect an observable frequency shift when changing from a high vowel to a low vowel. This is consistent with the assumption that $p(t)$ is not used to compensate for $s(t)$ – i.e. that they are independent.[5]

II. A TOY SUPERPOSITION MODEL

I will illustrate the behaviour of the model with simple idealized forms for $s(t)$ and $p(t)$. This section explores the mathematical consequences of superposition models of intonation, and is not proposing particular forms or numerical values for any particular sentence.[6]

In this example, the equation for $s(t)$ is chosen to represent an alternating sequence of phones that have intrinsic frequency shifts that differ by 8 Hz in successive syllables. The segmental part is thus taken to be

$$s(t) = 4 \cdot \sin(2\pi t/d) \tag{2}$$

(in Hertz), where $d = 0.5$ seconds is the period after which the segmental structure repeats; successive syllables can be imagined to occur every $d/2 = 0.25$ seconds.

The intonation part is

$$p(t) = 170 + 29 \cdot \cos(2\pi(t - \tau)/D), \tag{3}$$

which is a wave with period $D = 1.3$ seconds (i.e. about 5 syllables). Low and high intonation targets are separated here by two or three syllables. In this example, the average $f_0 = 170$ Hz is chosen to be midway between typical male and female values. The accents cause plus or minus 29 Hz pitch excursions (roughly 3 semitones) relative to the average. Figure 1 shows the underlying intonation contour $p(t)$, as a dashed curve and the surface f_0 contour, $f(t)$. The segmental shifts are small compared to the assumed intonation.

The time at which the first peak in $p(t)$ occurs is given by τ in Equation 3. Figure 2 displays a set of different $p(t)$ curves, each shifted slightly. These shifts correspond to changes in the alignment of the intonation peaks relative to the segmental structure of the sentence.

Figure 2 shows the intonation contours generated by the model. They differ only in the value of τ, which increases from $\tau = -50$ milliseconds in steps of 30 milliseconds. This set of curves corresponds to changes in the alignment of the intonation component $p(t)$ relative to the fixed segmental structure $s(t)$. These curves are simply shifted versions of the dashed curve in Figure 1.

This example uses frequency shifts caused by vowels, however, consonants can also contribute substantial shifts to $s(t)$. If there were only shifts on vowels, then peaks or valleys of $s(t)$ would be aligned with the syllable centers. But, if one includes segmentally specified f_0 shifts from consonants, peaks and valleys of $s(t)$ can also occur between syllables or at the edges of syllables. In this toy example, I plot curves in terms of a shift in $p(t)$, but equivalent results could be obtained by holding $p(t)$ constant and changing the pattern of segmental shifts. This is because the mathematics depends only on the alignment *difference* between segmental and prosodic peaks.

Figure 3 shows the resulting f_0 contours produced from the curves in Figure 2 by way of Equations 2 and 1. As the intonation contour shifts with respect to the segments, a sudden jump in the time of the maximum occurs between $\tau = 70$ milliseconds and

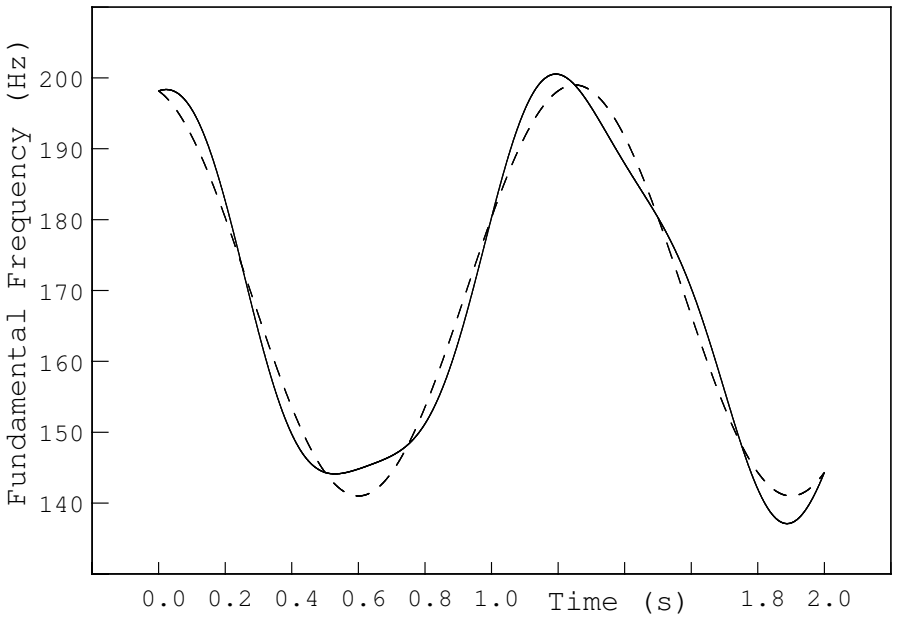

FIG. 1. The underlying intonation contour $p(t)$ (dashed) and the surface intonation $f(t)$ (solid line) for the model utterance. The plots shows fundamental frequency vs. time. The difference between the contours is the segmental effect, $s(t)$.

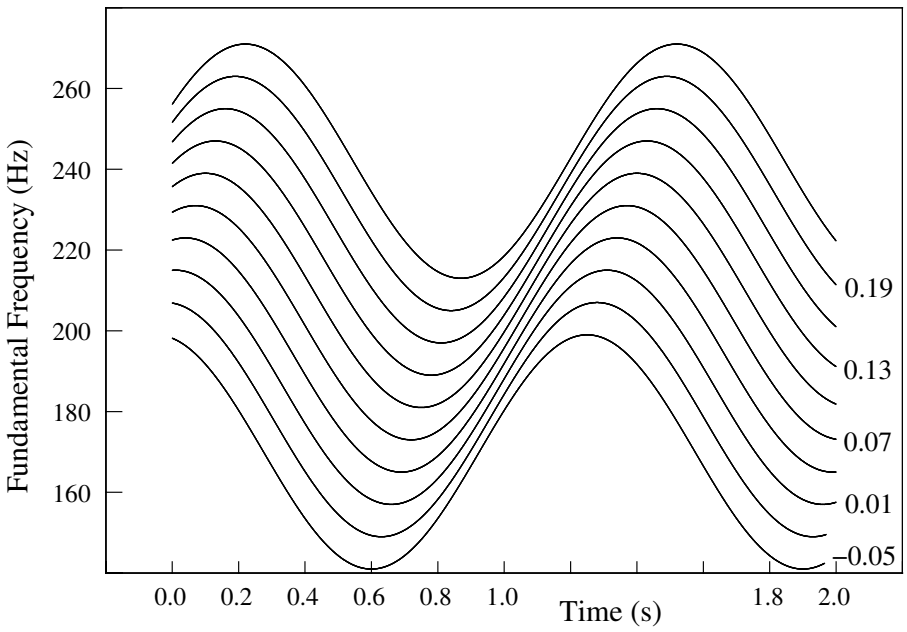

FIG. 2. A set of contours for the intonation component, $p(t)$. These form inputs to the intonation model, Equation 1. The contours are shifted vertically for clarity. Contours are labelled by τ (in seconds) and the maxima of neighboring curves differ by 30 ms.

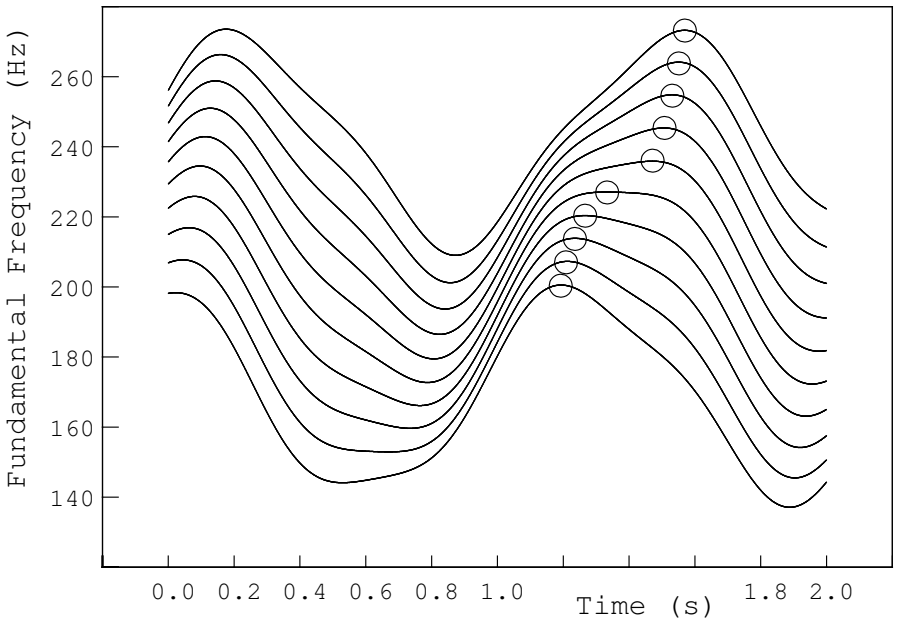

FIG. 3. A set of contours for $f(t)$, derived from Figure 2 by adding segmental effects (Equation 2). Small circles mark the maxima of the contours. Plotted as per Figure 2.

$\tau = 100$ milliseconds.

It is perhaps surprising that segmental effects which shift f_0 up and down can lead to substantial differences between the timing of the peak of the underlying prosodic contour and the surface f_0 contour. Rather than being on top of peaks in the underlying prosodic contour, peaks in f_0 are pulled towards nearby segments that cause a positive shift in f_0.

This can result in segmental anchoring: Imagine a group of utterances that have the same segmental structure but progressively different underlying prosodic f_0 contours. In such a situation, an f_0 peak can be "pinned" to a particular segment that has a positive segmental effect. Even if the peak of $p(t)$ shifts from utterance to utterance (within some range), the peak of f_0 might always occur within the same segment.

If there are two nearby segments with positive segmental effects with the underlying peak of $p(t)$ in between, the f_0 peak can switch from being pinned on one to being pinned on the other. This can lead to an effect where the f_0 peak jumps a substantial interval (e.g. the spacing between two syllables) as the result of an arbitrarily small change in the alignment of the underlying prosodic contour. Such a jump could easily be misinterpreted as evidence for distinct phonological states of prosodic alignment.

A. Explaining the Jump

One can understand the sudden jump by considering two limiting cases of the model. First, suppose that the amplitude of $s(t)$ is very small compared to $p(t)$. Then, the time of the observed f_0 peak would smoothly follow the peak of $p(t)$.

Now, consider the opposite case where $p(t)$ is small compared to $s(t)$. Then, segmental shifts would dominate, and the overall maximum would always be very near one of the maxima of $s(t)$, but the underlying prosody would select which maximum would be the highest. One of the maxima of $s(t)$ would be pushed up by the (tiny but nonzero) $p(t)$. Therefore, if one were to vary τ, the maximum of $f(t)$ would move only occasionally, when it would jump from one maximum of $s(t)$ to the next. (This case describes an extreme form

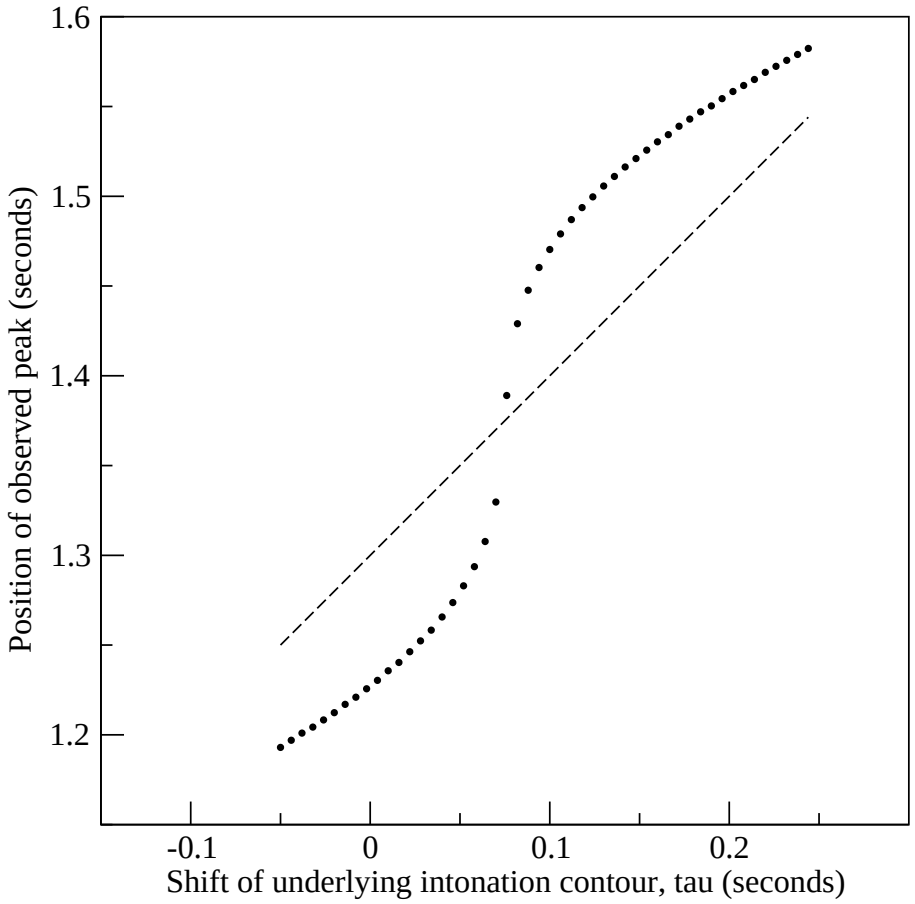

FIG. 4. The relationship between the time shift τ of the underlying intonation contour and the maximum of the observed fundamental frequency, argmax$\{f(t)\}$. The vertical axis indicates the position of the maxima (open circles) in Figure 3. The diagonal dashed line corresponds to the position of the underlying prosodic peak.

of segmental anchoring, where strong correlations arise between measured f_0 maxima and segment locations.)

For realistic values of $s(t)$ and $p(t)$, the model gives results intermediate between these two extreme cases. The maximum of $f(t)$ is loosely anchored to the maximum of $p(t)$, but it changes non-uniformly as the underlying alignment changes.

Figure 4 plots the relationship between the maxima of $f(t)$ and τ. (It is a more detailed plot of the times of the maximum at the left edge of Figure 3, near $t = 0$.) It is s-shaped

and strongly nonlinear. There is a non-trivial relationship between the time of the observed peaks in $f(t)$ and the time-alignment of the accents, which is the underlying quantity of interest to intonational phonologists.

The nonlinearity becomes more pronounced as the magnitude of $s(t)$ increases or as the magnitude of $p(t)$ decreases. In fact, if the curvature (i.e. the second time derivative) of $s(t)$ were to exceed the curvature of $p(t)$, Figure 4 would have a discontinuous jump so that an infinitesimal change in τ would lead to a substantial jump in the time of the observed f_0 peak. These jumps occur when the top of the maximum is flat (or nearly so); such flat-topped f_0 contours are described in Knight (2002), Ogden *et al.* (2000, especially pp. 194–195, Figure 8) and others. With a small pitch range in $p(t)$ or flat-topped profile, a small shift in alignment or segmental structure can even move the time of peak f_0 from one syllable to the next.

B. Segmental Anchoring

Segmental anchoring occurs when a peak in $s(t)$ approximately coincides with a peak in $p(t)$. This should happen when high vowels or certain consonants are near accents that raise f_0. The mechanism is discussed in §II.A.

The anchoring effect can be important even in tightly controlled experiments that generate several intonations for the same text. Dilley *et al.* (2005) test whether accents are anchored to each other or to the segmental structure, and conclude that for their data, the anchoring is segmental. Several other papers can be interpreted as showing anchoring of f_0 contours to the syllable (Arvaniti *et al.*, 1998; van Santen and Hirschberg, 1994; Ladd *et al.*, 2000), possibly because of this effect.

C. False Phonological Contrasts

The opposite of segmental anchoring occurs when a prosodic peak approximately coincides with a segment that pushes f_0 down. This is the situation shown in §II.

The resulting nonlinearity shown in Figure 4 is important because it could lead a researcher to falsely assert a discrete (and thus presumably phonological) contrast when none really exists. Suppose that the alignment of $p(t)$ is somewhat variable and that the alignment of the peak is taken from a single unimodal distribution with no phonological distinctions. For example, take τ from a Gaussian distribution with a mean of 50 milliseconds and a standard deviation of 40 milliseconds. One can think of this as a corpus containing many utterances, each with a slightly different alignment.

To show the effect, the average peak alignment will be placed midway between maxima of the segmental effect. Figure 5 shows that despite the unimodal distribution of alignments, the distribution of observed f_0 maxima is bimodal.[7] This could easily lead to a false belief that there are two underlying phonological categories (e.g. early peak vs. late peak), one for each maximum.[8][9] In reality, though, the bimodal distribution is generated by an interaction between the intonation, the segmental shifts and the measurement procedure.

Thus, taking the maximum of an f_0 contour and interpreting it as an intonation peak is a dangerous procedure. Small segmental effects can lead to large changes in the position of the maximum. The root cause of the trap is that the peaks of observed f_0 that one *can* measure are not the same as peaks in the prosodic contour, $p(t)$, that one would *like* to measure. An algorithm is therefore needed to estimate peaks in $p(t)$ from observations of $f(t)$.

D. The Bracketed maximum Algorithm

It has been shown that the simple maximum[10] of $f(t)$ is not a satisfactory estimate of the maximum of the underlying prosody. But how can one do better? Ideally, one would model the segmental effects $s(t)$ and simply subtract that function from $f(t)$ to directly yield the desired $p(t)$. However, this would not be trivial, as segmental shifts seem to vary from speaker to speaker. Intonation experiments and data analysis would have to become rather larger and more complex to estimate the segmental shifts of each phone for each speaker in

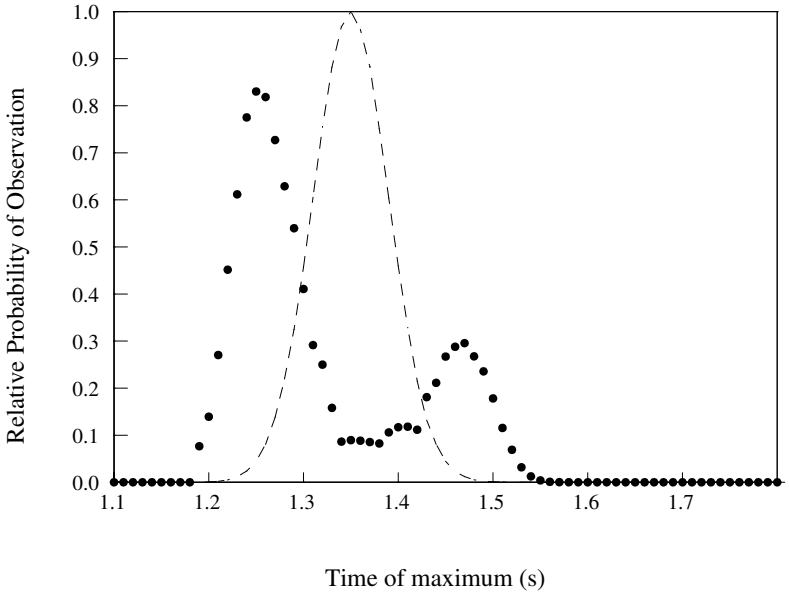

FIG. 5. The distribution of the times of f_0 peaks including segmental shifts (dots). This histogram assumes a Gaussian distribution of τ (the location of the underlying intonation maxima) with a standard deviation of 40 milliseconds. (A shifted version of the distribution of τ is shown by the thin dashed line.) The horizontal axis is the position of the maximum and the vertical axis is the probability of observation, collected into 10 millisecond bins.

all relevant contexts.

A more practical approach is the "bracketed maximum",[11] which springs from two ideas. First, if one thinks of the segmental effects as "noise", one might be able to get a better measurement of the prosodic peak position if we could take the average of two measurements from locations where the segmental effects were approximately independent.[12] This is possible by measuring in two different segments.

Normally, peaks in the prosodic component of f_0 are expected to be wide enough to cover more than one phoneme and their left and right edges will often be in different phonemes with different, almost uncorrelated segmental effects. As long as the two measurements are close enough to be on the same accent, it is plausible that the errors in the average position may be smaller than the error in either individual measurement.

Second, making a timing measurement on the side of a peak is intrinsically less sensitive to segmental effects than a measurement near the top of a peak. Suppose one is looking for the top of a broad maximum. There might be a region perhaps 50-100 ms long over which the frequency changes by only 4 Hz. Now, if segmental effects were different, so that f_0 changed by a few Hertz, then the observed maximum might move anywhere within that large region, controlled by the pattern of segments.

On the other hand, if one makes timing measurements on the side of a peak, they will be less affected by segmental effects because f_0 is changing rapidly. To take a plausible example, if $p(t)$ swings through 40 Hz in 100 milliseconds, it will take only 10 ms to swing through the 4 Hz range corresponding to segmental effects. So, if the segmental structure of the sentence were rearranged without changing $p(t)$, one would expect the time at which f_0 crosses a given threshold to change by only 10 milliseconds or so. Changing the segmental effects will thus have a smaller effect on the timing measurements if the measurements are made on the sides of the peak.

The bracketed maximum is a simple technique that implements these ideas and can reduce the effect of segmentally determined frequency shifts. It involves measuring on both sides of the maximum of $f(t)$ and averaging the two measurements. It should work as long as

the segmental shifts are small compared to the prosodic f_0 swings. (Edge cases are discussed in Appendix B.)

The technique involves four steps

- Find the maximum of f_0.
- Go backwards from the maximum as long as f_0 is within Δ Hz of the maximum, or until an unvoiced region is encountered. Call this time t_L. Here, Δ is called the measurement offset.
- Go forwards from the maximum, as long as f_0 is within Δ Hz of the maximum,[13] or until you encounter an unvoiced region. Call this time t_R.
- Average t_L and t_R to produce an estimate of the time at which $p(t)$ is maximal. This is an estimate of the underlying intonation peak alignment.

Note that for $\Delta = 0$, the result is just the simple maximum. The algorithm may be downloaded from `http://kochanski.org/gpk/papers/2008/Segmental_Additive/algorithm.py.txt` and/or as part of this paper's supplemental materials.

(Figure 6 shows sample data annotated to show the operation of the bracketed maximum algorithm.)

In the toy model of §II, this algorithm yields a more accurate and more linear relationship between the observations and τ, compared to the simple maximum. Figure 7 shows a comparison between τ the result of the bracketed maximum algorithm and the result of the simple maximum.

Another advantage of the bracketed maximum can be seen in Figure 8. This is computed as per Figure 5, except that it shows the distribution of peak positions for both the bracketed maximum and simple maximum algorithms. The bracketed maximum gives a unimodal distribution and thus does not falsely suggest a phonological distinction between two categories. The full-width at half-maximum is 107 milliseconds, close to the 94 millisecond full-width at half-maximum for the distribution of τ.

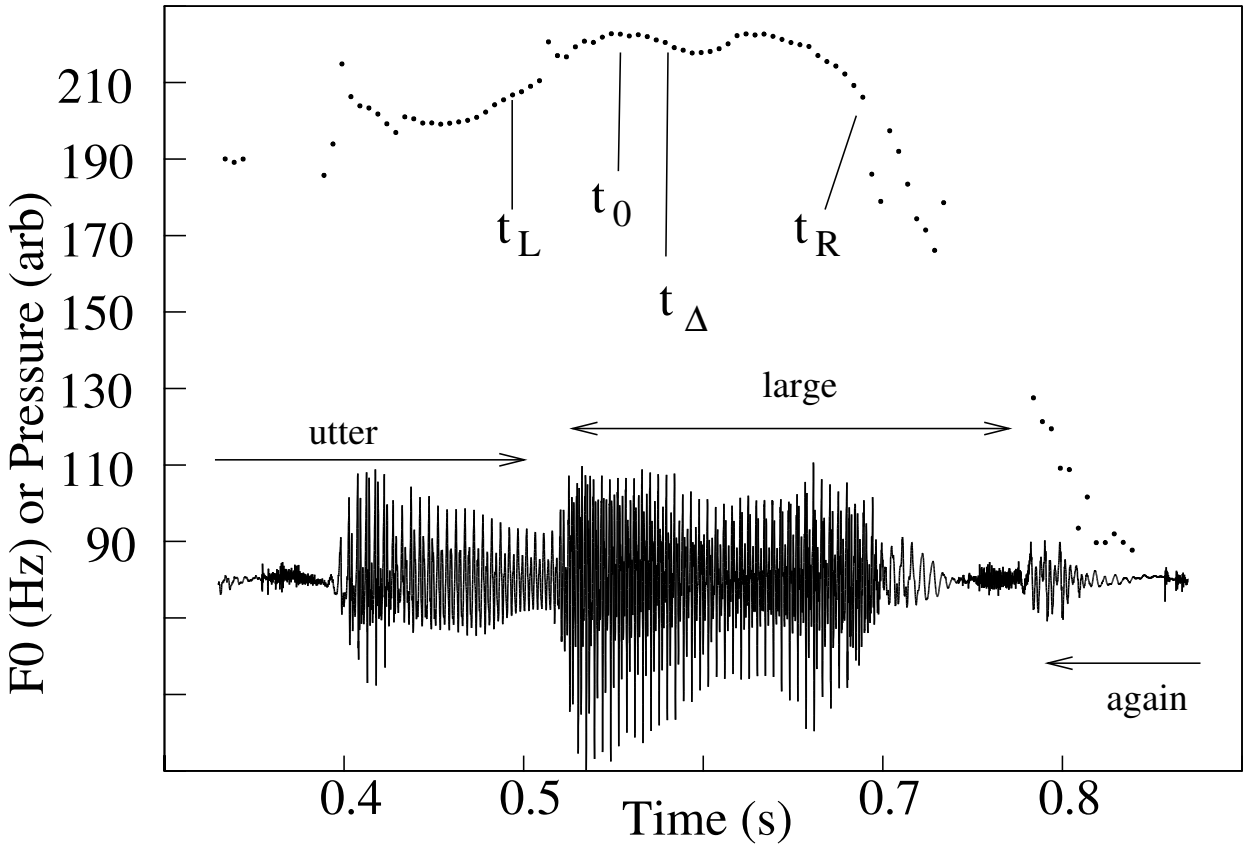

FIG. 6. Sample audio data is plotted (lower curve) with a transcription, and an f_0 contour above that. The f_0 contour is annotated with the simple maximum (labeled t_0), intermediate results from the bracketed maximum algorithm with $\Delta = 20$ Hz (labeled t_L and t_R), and the final result (labeled t_Δ). This data was chosen to show the operation of the bracketed maximum algorithm on a broad f_0 peak.

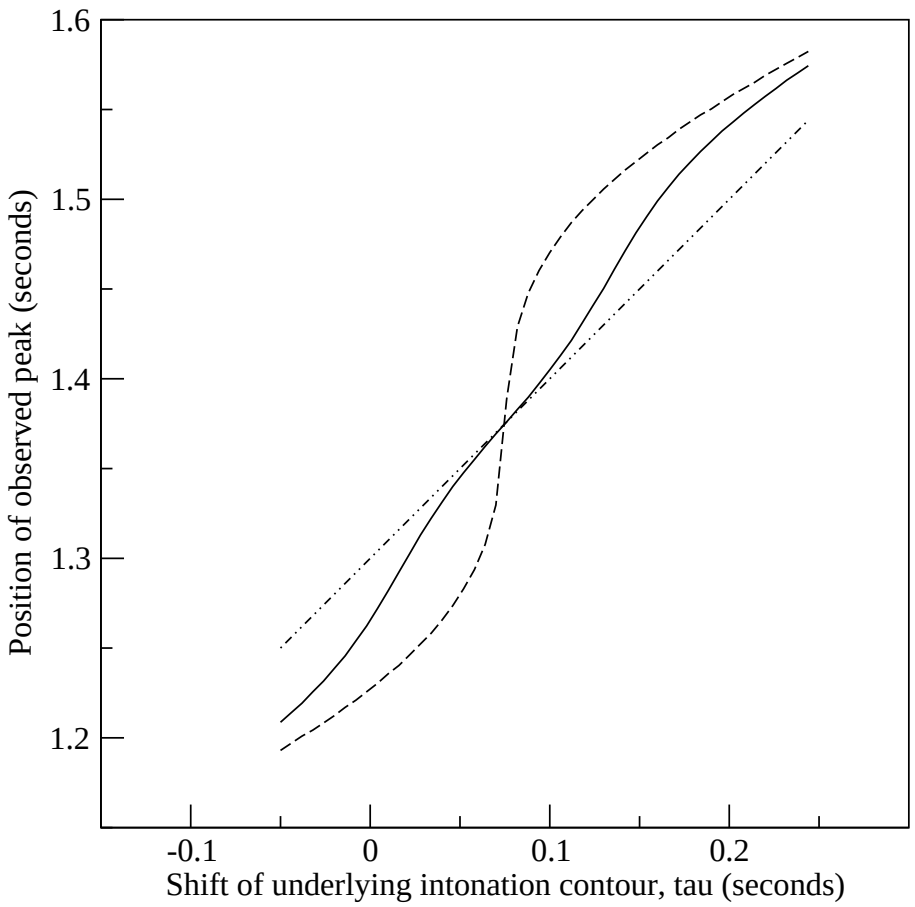

FIG. 7. The bracketed maximum (solid), the simple maximum (dashed) *vs.* τ, the maximum of the underlying intonation contour.

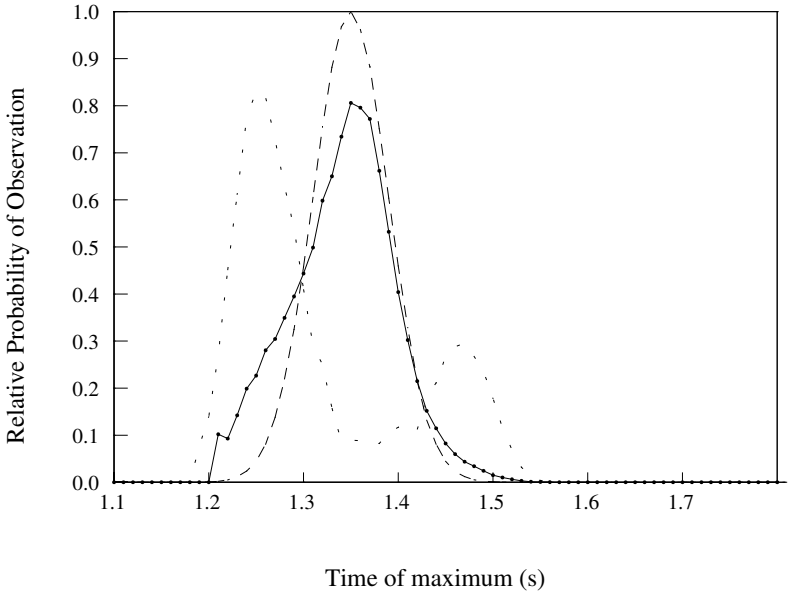

FIG. 8. The distribution of estimated intonation peak positions, $(t_L + t_R)/2$, for a simulated experiment using the bracketed maximum technique (solid line w/dots). The resulting histogram is unimodal, reflecting the unimodal distribution of τ (alignment). Results from the simple maximum (from Figure 5) are reproduced as a dotted line for comparison, and the underlying distribution of the prosodic peak position (τ) is shown as a dashed line.

The measurement offset (Δ) that one uses is not particularly critical in this "toy" model. Performance gradually improves as the measurement offset is increased, with a broad optimum when it is half of the peak-to-peak swing of $p(t)$ (e.g. 29 Hz in this example). However, the bulk of the improvement comes by the time the measurement offset reaches the peak-to-peak segmental effect, $s(t)$ (e.g. 8 Hz).

E. Limits on the choice of Δ

One limit on using the bracketed maximum algorithm on real speech is that one does not know the size of the relevant pitch excursions. Since the bracketed maximum will not produce a useful value for f_0 maxima that are smaller than the measurement offset, Δ should not be made too large.

However, very small f_0 maxima are unlikely to be perceptually important or phonologically distinctive. When it is important for the speaker to communicate a distinction between two possible meanings of a sentence, one expects that he or she will produce an easily perceptible peak, large enough to be reliably detected by the listener. Under favorable conditions, frequency differences as small as about 3 Hz may be detectable (Chuang and Wang, 1978), but other experiments show that larger pitch motions are needed (Peng, 2000, §10.3). Mack and Gold (1984) showed that the minimum detectable pitch shift is a function of the complexity of the stimuli, ranging from 2 Hz for a buzz-tone though 4 Hz for monotone sentences, to more than 6 Hz for sentences with near-natural intonation patterns.

Since the discrimination threshold is typically defined as a 75% correct detection of a difference under quiet laboratory conditions without distractions, one expects that a yet larger shift is necessary for reliable communication in realistic conditions.[14]

Linguistic evidence also points to somewhat larger values. For instance, Holm and Bailly (2002, §3.1) note that different repetitions of the same utterance typically differ by 1 semitone (about 10 Hz). Braun *et al.* (2006, §III.C) obtained similar results, suggesting that phonologically distinct utterances are separated by about 3 semitones, while changes

smaller than about 1 semitone were unimportant.

Overall, these considerations suggest that any phonological differences that the speaker wishes to be understood by the listener will probably be encoded by pitch shifts of 10 Hz or larger. So, with Δ near 10 Hz, the algorithm should produce reliable results for most f_0 maxima that have any communicative function.

III. TESTS ON REAL SPEECH

A. Introduction

The argument so far rests on the basis of an idealized mathematical model (although the parameters used in the model are consistent with observations). One of the predictions of the model is that given a corpus of sentences with the same prosodic contour but different segmental structures, segmental effects would make the timing of the observed f_0 peak vary.

Thus, a measurement technique like the bracketed maximum should reduce the variability of peak timing in such a corpus. We can check this chain of logic experimentally by comparing various measurement algorithms. Whichever procedure produces the smallest variance in peak timing is presumably least affected by segmental effects.

This simple maximum algorithm will now be compared to

- The Bracketed Maximum, described in §II.D.
- Smoothing the data with a median filter, then taking the time at which the median is maximum. This follows Taylor, 1993 and Xu and Xu, 2003. It ignores variations of f_0 on short time scales such as 100 milliseconds or less. These short time scales are where the segmental effects are most dramatic.
- Smoothing the data by averaging over a window centered around each point, and then taking the time when the average is maximal.
- The MOMEL algorithm of Di Cristo and Hirst (1986); it transforms intonation contours into a quadratic spline approximation, producing a smooth representation

that bridges over unvoiced regions.

A priori, one expects that these approaches generally will smooth away structures on short time scales (e.g. segmental effects) and will generally preserve longer, suprasegmental f_0 motions.

B. Data

I use a corpus of utterances with fixed sentence patterns where there is reason to believe that the intonational phonology always specifies a peak in the same position. One can then compare algorithms to see which one gives a more stable estimate of the intonation peak.

The database consists of single-syllable words embedded within a frame sentence (Slater and Coleman, 1996); it was previously collected for a different purpose. The corpus contains 4970 utterances from a single speaker of Southern British English, each recorded five times. There are four frames used: "Can you utter 'X' again, please?" (used where X is a word in the form CxC, beginning and ending with a consonant), "Can you utter 'X' today?" (X is CxV, beginning with a consonant and ending with a vowel), "Have you uttered 'X' again?" (X is a VxC word), and "Have you uttered 'X' today?" (X is VxV). The speaker was phonetically trained and knew he was reading a list containing minimal pairs of English words. In such a database, the frame sentence and the words are semantically neutral, and the syntax is such that any word can be used as X. Consequently, the intonation should be identical within groups of utterances that share the same frame.

One expects an accent on the variable word, because it is most informative. (General information on focus and accent location can be found in Ladd (1996, §5.1) and references therein.) The accented word should thus be louder and longer than its neighbors (e.g. Kochanski *et al.*, 2005 and references therein). A random sample of 100 utterances were checked by the author, and an obvious prominence was heard on the variable word in 98 cases. In three of those 98 cases, words in the frame were judged to be as prominent as the variable word. So, in 95 out of the 100 samples, the variable word was the most

prominent word in the utterance.

C. Signal Processing and Overview of Data

The acoustic data were processed to extract time-series measurements of quasi-duration, loudness[a], f_0, and aperiodicity, as per Kochanski *et al.* (2005). Time axes were normalized to span the range from 0 to 1 between the beginning of the second syllable of "utter"/"uttered" and the end of the first syllable in "today"/"again". Thus, normalized time 0 to 1 always spans three syllables and the variable syllable is centered near 0.5.

For each utterance, we compute estimates of the prosodic peak position for different values of the measurement offset Δ. This yields values $t_i(\Delta)$ where i indexes the utterance.

We included only f_0 maxima on or near X by restricting the analysis to f_0 data with maxima at normalized time between 0.2 and 0.8. We did this because the f_0 in "utter"/"uttered" can sometimes be higher than the peak f_0 within the variable syllable[15] (see Figure 11); this can lead to unexpected results in the computed peak positions. Utterances where either t_L or t_R was outside that region were dropped.

A scatter-plot of the quasi-duration for the corpus is shown in Figure 9; this is computed per Kochanski *et al.* (2005). (At each time, the quasi-duration measures how far one can go forward and backwards in time before the spectrum changes substantially. The quasi-duration at any time is roughly proportional to the duration of the phone at that time.) The frame syllables are centered near normalized times of 0.17 and 0.82 and the variable syllable is approximately centered. The variable syllable typically has a longer vowel with a more stable formant structure than found in the frame syllables. (The region near normalized time 0.3 corresponds approximately to the boundary between the preceding frame syllable and the variable syllable.) A plot of an estimate of the perceptual loudness appears in Figure 10. The variable syllable can be seen to be typically longer and louder than its neighbors, and thus should typically be prominent. (Incidentally, one can see a bimodal distribution of loudness from the two possible frames near normalized time 0.25.)

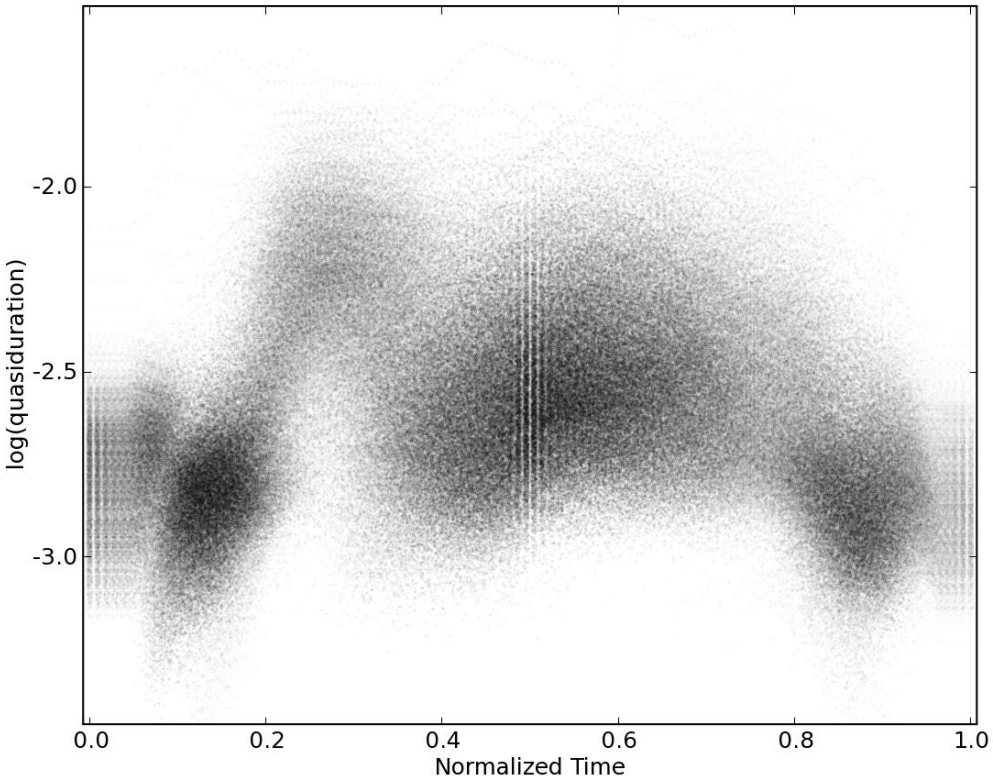

FIG. 9. Quasi-duration scatter-plot for the entire corpus. The quasi-duration is a measure of the stationarity of the speech spectrum; small values imply a rapidly changing spectrum. The plot shows the log of the normalized quasi-duration against normalized time, otherwise plotted as per Figure 11.

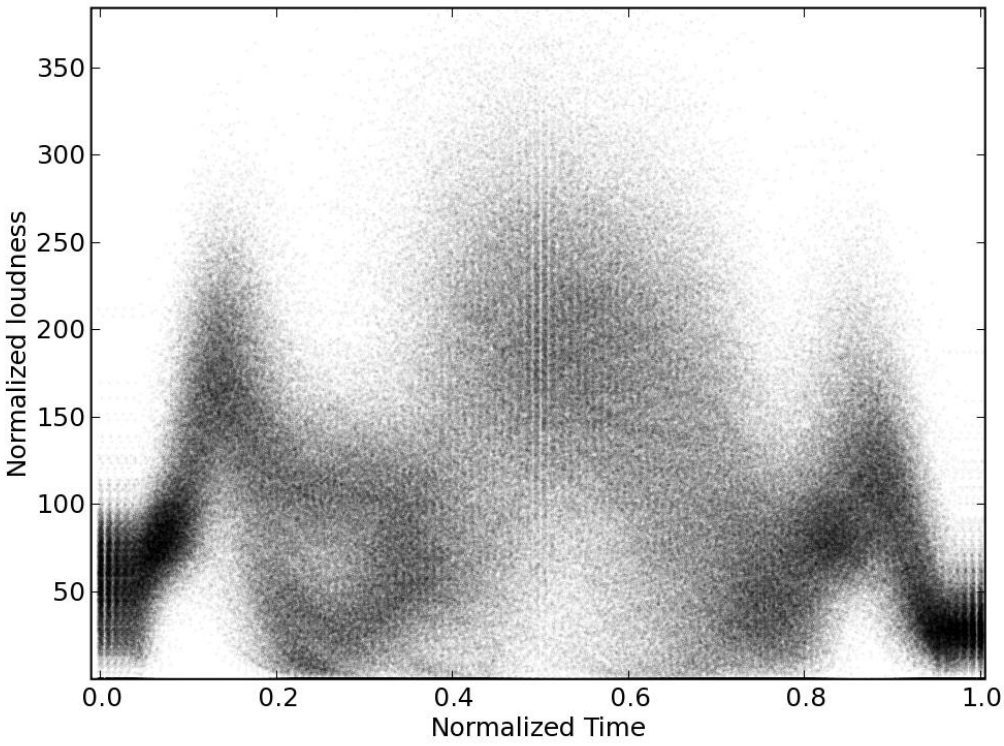

FIG. 10. Normalized loudness vs. normalized time for the entire corpus. Plotted as Figure 11, except that all dots are the same size.

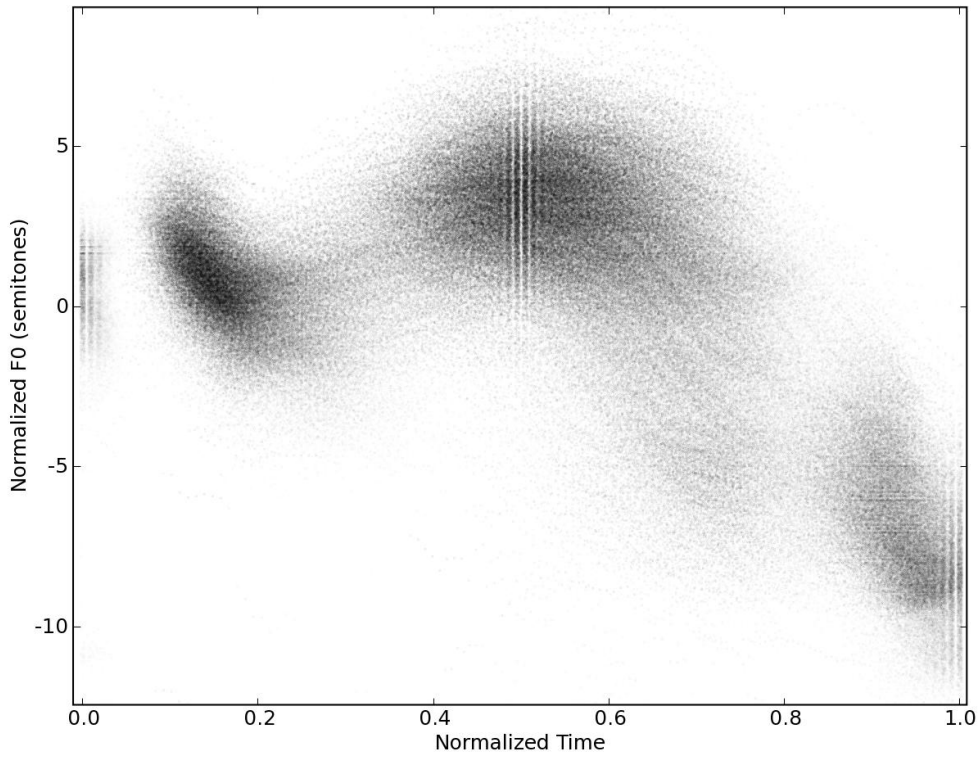

FIG. 11. Fundamental frequency of the entire corpus. This plots normalized f_0 against normalized time for the region of interest. The horizontal axis goes from the beginning of the syllable before the variable syllable to the end of the syllable after. The vertical axis is f_0 deviation from 170 Hz, in semitones. This is a smoothed scatterplot of f_0 measurements. (The vertical stripes, e.g. near $x = 0.5$ are the result of the 10 ms interval between f_0 measurements.)

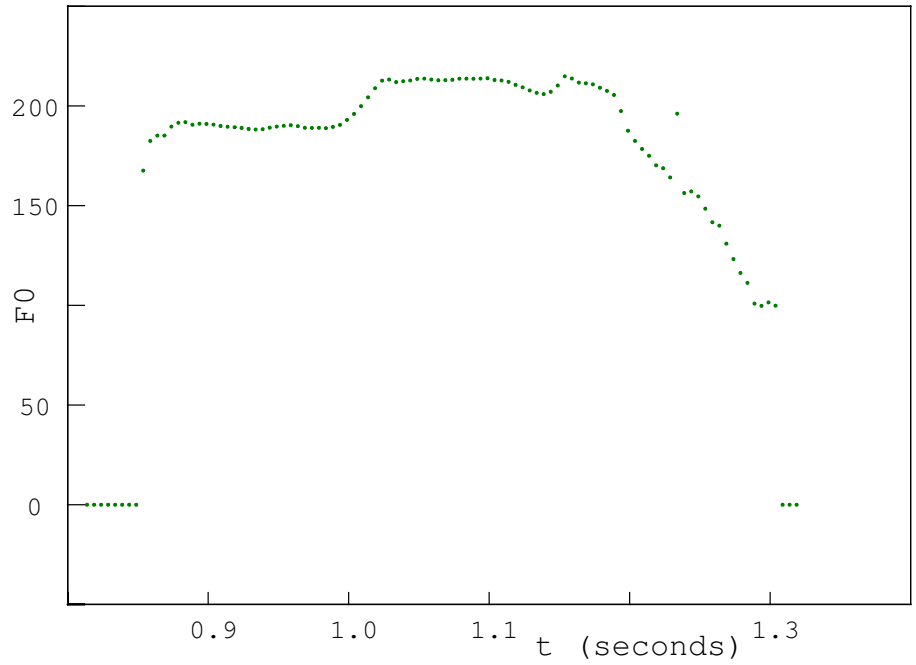

FIG. 12. Sample flat-topped f_0 profile.

Figure 11 gives an overview of f_0 data from the entire corpus. It shows a three-syllable region centered on X, plotted as per Braun *et al.* (2006). It is created by placing 342000 dots, one per f_0 measurement. In the plot, dot sizes are proportional to an estimate of the perceptual loudness in regions where the speech waveform is approximately periodic and are reduced where the waveform is aperiodic. This emphasizes loud, periodic regions (e.g. the central vowel of each syllable and regions where the f_0 data is most reliable). The image has been smoothed slightly to reduce printing artifacts; this blurs the individual dots slightly. The variable syllable typically also has an f_0 peak, providing further evidence of its prosodic prominence.

1. Flat-Topped Profiles

In many utterances, the f_0 peak position is not completely clear, usually because the f_0 curve has a flat top with no obvious peak. These are commonly known as Plateaus (House *et al.*, 1999; Ogden *et al.*, 2000; Wichmann *et al.*, 1997). Figure 12 shows one such example.

(Note that this utterance was chosen to display a plateau; for f_0 data representative of the corpus as a whole, see Figure 11.)

To count flat-topped utterances, I looked for utterances where many points would likely to be indistinguishably high. I defined top points to be voiced points within 3 Hz of the second-highest f_0 value. (The point with highest f_0 was ignored on the grounds that it was commonly at the edge of a voiced region and the waveform was usually not stationary there.) The f_0 differences among these top points are small enough so that the listener is unlikely to be able to reliably pick one as higher than another. (See §II.E.) Considering the speaker, small muscle tremors or other linguistically unimportant changes in production could push f_0 up or down by 3 Hz, so any top point could plausibly have been an intended intonation maximum (this follows House *et al.*, 1999; Knight and Nolan, 2006).

In each utterance in the corpus, a computer program found the largest interval starting and ending on top points that contains at least 50% of top points. The length of this interval provides an estimate of the uncertainty in the time of the maximum: it measures the length of the flat top of the utterance. Using this criterion, 9% of the utterances have intervals of plausible maxima longer than 100 ms and 36% have intervals longer than 60 ms. This criterion, while somewhat arbitrary, shows that a substantial number of utterances have tops that are flat enough so that segmental f_0 perturbations could lead to large changes in the timing of the maximum.

D. Results: Comparison of Algorithms

To compare different approaches, we computed the variance of the peak position over the corpus. Since the underlying accent is expected to be in a stable location, the algorithm that provides the most stable measurement should be the most accurate. (See Appendix A for a justification of this assumption.) This idea is a common way of choosing a measurement technique; see Kochanski and Orphanidou, 2008 in phonetics and Sachs *et al.*, 1995, Krolik, 1996, Abel, 1990 in other fields. However, this paper is not a full-blown competitive

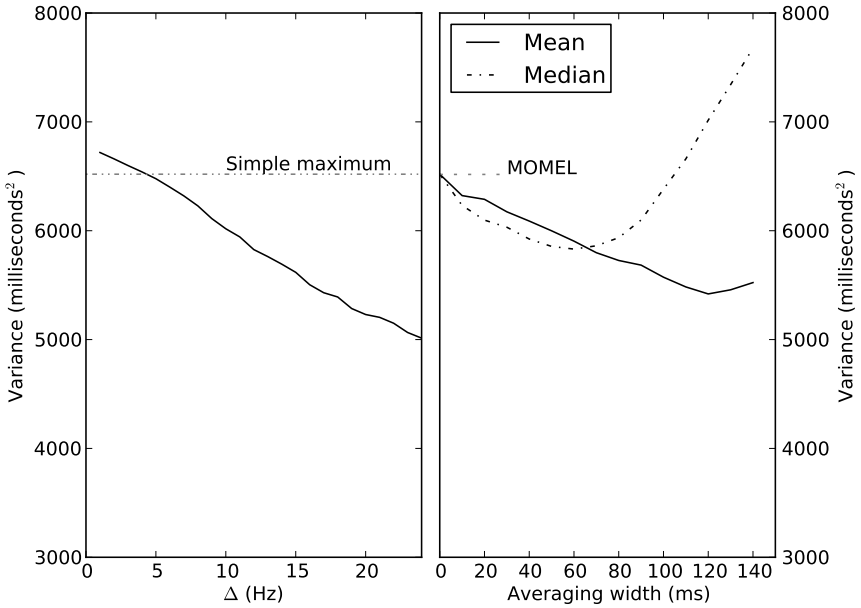

FIG. 13. Left: variance of the bracketed maximum estimate of the prosodic peak position, plotted against Δ. The solid line shows the variance of the entire corpus. The simple maximum is shown by the dash-dot line, for comparison. Right: variance of algorithms that smooth the f_0 data and then take the maximum. The variance is plotted against the width of the smoothing window. The best MOMEL result is the grey dot. (MOMEL does not have a comparable smoothing parameter, so the horizontal axis is irrelevant.)

evaluation, primarily because the testing corpus is produced by a single speaker.

In our corpus, there is a built-in dependence of the frame sentence on the initial or final segments of the variable word. This could possibly cause the prosodic peak position to depend on the segmental structure if the type or alignment of the accent depended on the choice of frame. However, this problem is avoided by splitting the corpus into four sub-corpora, one for each possible frame, and then computing the variances separately within each sub-corpus.

Figure 13 shows this variance for different algorithms. (See Appendix C for comparisons of the various sections of the corpus.)

As the measurement offset is increased, the variance of the bracketed maximum progressively decreases. Thus, the best measurement of the underlying accent position is made with Δ as large as possible. For all $\Delta > 3$ Hz in this data set, the variance is reduced by a statistically significant amount. The significance level improves from $P < 0.05$ at $\Delta = 3$ Hz to $P < 10^{-6}$ at $\Delta = 10$ Hz. For $\Delta = 10$ Hz, the variance is reduced by 9%, and for $\Delta = 20$ Hz, the variance is reduced by 21%.

Note that this reduction is a fraction of the total variance, which includes both the errors induced by segmental effects and also the intrinsic variability in the peak alignment (see Appendix A for discussion). Thus, the fractional reduction quoted here gives a lower bound to the reduction of the segmentally-induced measurement error.

The largest practical Δ is set because the algorithm can only measure peak positions if the peaks are larger than Δ. For this dataset the the fraction of unmeasurable syllables is small (less than 1%) as long as $\Delta \leq 20$ Hz. This is consistent with §II.E because the psychophysical and linguistic evidence can only establish minimum peak sizes for intelligibility; nothing prohibits a speaker from using an f_0 peak larger than the minimum. However, this 20 Hz upper limit for Δ should be treated as an estimate, as substantial inter-subject and inter-style differences in f_0 range are not unknown.

When this new procedure is used, the largest possible value of Δ should be used, so long as linguistically important f_0 peaks are not lost. Users should be aware of the tradeoff between large Δ, which provides the most reliable timing measurements but cannot measure smaller peaks, vs. small Δ, which provides less advantage over a simple maximum but can operate on small peaks.

The right side of Figure 13 shows corresponding results for two algorithms that smooth the f_0 contour and then take the simple maximum. The performance of these algorithms is plotted as a function of the width of the smoothing window. Using an arithmetic smooth with a 110 ms-wide window (i.e. taking the mean of all points that are voiced and within 55 ms of the point under consideration) reduces the variance by 13%.

However, the variance worsens for large smoothing windows. Presumably, the smoothing

window is becoming wide enough to significantly distort the shape of $p(t)$. This optimum point may be the width where the increasing systematic distortion becomes more important than further reductions in segmental effects. Since this optimum point is likely to depend on the speech rate, mean spacing between accents and other characteristics of the corpus, in practice, it would be safer to use a smaller smoothing window. With a 90 ms window, an arithmetic smooth improves the variance by 12%. This improvement is very similar to the bracketed maximum algorithm's performance when operated with a relatively safe value of $\Delta = 10$ Hz.

Results for the median smoothing algorithm are similar, but not as good (for the particular speaker who produced our corpus). The optimal window width is then 60 ms, which yields an improvement of 10% in the variance. The median smooth behaves very badly for large window widths; for windows 100 ms or wider, it is worse than the simple maximum. This is doubtless related to the flat tops that are typically produced on intonation peaks when $f(t)$ is subjected to a median smooth. After a median smoothing operation, there is often no unique maximum, with several points near the top of the peak having mathematically identical values. The final step of applying the simple maximum to the smoothed curve then can behave badly.

Both of these smoothing-based algorithms can improve the accuracy of the timing estimation, but the bracketed maximum can out-perform them under some conditions (i.e. for this corpus when $\Delta > 14$ Hz).

MOMEL (Hirst and Espesser, 1993; Di Cristo and Hirst, 1986) does not yield an improvement to the peak location accuracy. We ran MOMEL in 180 different ways; the best performance in terms of timing variance is plotted in Figure 13 (right), which is nearly the default (Hirst and Espesser, 1993) parameters. All the runs used f_0 data sampled at 10 ms intervals. The best-performing six sets of parameters are all close to the recommended defaults for the program, with $1.032 \leq \text{maxerr} \leq 1.044$, and other parameters falling between 80% and 110% of their default values of win1 = 30, win2 = 20, mind = 5, and minr = 0.05; none of the top six runs used the "–non-elim-glitch" flag. All of the top 45 runs (including the

run plotted) mask MOMEL's output with a voicing indicator, so that it does not interpolate f_0 into unvoiced regions. The alternative (treating MOMEL's result as valid in unvoiced regions) leads to a peak position variance which is 20% larger than applying either MOMEL or the simple maximum over voiced regions.

This result is perhaps not surprising, because MOMEL is used differently from its intended application. It was intended to represent entire utterances, and intended to provide an smooth f_0 curve that is intonationally equivalent to the input data. Here, it is being used on a small fragment of an utterance, extracted from the middle, and we are expecting it to smooth away segmental effects.

IV. CONCLUSION

Segmental effects are surprisingly important to experiments that measure the timing of f_0 peaks, if a superposition model of intonation is adopted. It has been shown that:

- peak positions can be strongly influenced by the segmental structure near the peak, and
- very small changes in the underlying intonation can lead to large jumps in the measured peak position.

The model shows that there can be unexpectedly large correlations of peak positions with the segmental content of the utterance. These problems arise because in a superposition model, the measurable quantity (f_0) is different from the underlying prosodic contour. Even with fairly large intonational swings (such as 3 semitones), segmental effects should not be ignored, and they are increasingly important when the pitch range becomes smaller. This effect can lead to the "anchoring" of f_0 peaks to segments that boost f_0. Conversely, peaks in f_0 will be repelled from segments with especially low f_0. Under plausible assumptions, this can generate a bimodal distribution of f_0 peak positions that is unrelated to any underlying phonological distinctions.

This work shows that any f_0 peak much less than 10 Hz tall will have its height and alignment strongly affected by segmental effects. Even if there is lexically identical comparison data, the peaks in f_0 will be systematically biassed towards segments with a higher intrinsic f_0. (Also, a review of the literature suggests that such a small peak is unlikely to reliably communicate anything to the listener.) Consequently, f_0 peaks with small excursions may best be ignored, or used only in situations where segmental effects are well understood. Overall, peak timing measurements need to be conducted with careful consideration of segmental effects.

These observations imply that peak timing measurements should not be simply interpreted as the alignment of the peak of an underlying prosodic contour in superposition models of intonation. Overall, segmentally-influenced differences between underlying and measured peak positions are large enough so that to properly interpret many experiments in terms of intonational phonology, a numerical model of segmental effects will be necessary.

Of course, superposition models are not mandatory. However, if one wishes to maintain the distinction between the surface f_0 and an underlying phonologically-determined intonation contour, some sort of model is necessary to connect from one to the other. It seems likely that similar effects will be seen for a broader class of models, including all those referenced herein. An alternative is that phonological analysis would be conducted on the observable f_0 directly. This would have the implication that f_0 should be identified directly with phonological intent, thus erasing any sort of competence/performance distinction.

This paper also presents a technique for measuring the timing of peaks, the "bracketed maximum." In simulations, it yields a substantially better estimate of the underlying prosodic peak position than simply taking the time of the maximum f_0. The bracketed maximum is tested on speech data, on a corpus where the accent positions are believed to be known à-priori. It can give significant reductions in the variance of the estimated peak position, and can behave as well or better than existing techniques that involve smoothing the f_0 contour.

V. ACKNOWLEDGEMENTS

I thank John Coleman, Elinor Payne, Burton Rosner, Chilin Shih and Christina Orphanidou for their valuable comments. John Coleman also kindly supplied the data set used here. I thank Daniel Hirst for release and explanations for the MOMEL code. I gratefully acknowledge financial support from the UK's Economic and Social Research Council under project RES-000-23-0149.

APPENDIX A: INDEPENDENCE AND MINIMUM VARIANCE

For a superposition model of intonation, it is possible to show that the best algorithm is the one that gives the least variance. Thus, one can compare two ways of estimating the underlying peak position by simply comparing the variance of the estimates.

Formally, we will compute the variance of the estimated peak positions from each algorithm. We consider the output of algorithm X, t_X, to be the sum of two random variables: the time at which $p(t)$ has a maximum (P) plus an error caused by segmental effects (S_X). If S_X is independent of P, the variances will add nicely so that

$$\mathrm{var}(t_X) = \mathrm{var}(P_X) + \mathrm{var}(S_X). \tag{A1}$$

Since we compare all the algorithms on the same corpus, $\mathrm{var}(S)$ is the same for all algorithms. Thus the algorithm with the smallest $\mathrm{var}(t_X)$ will have the smallest $\mathrm{var}(S_X)$, and therefore it will have the smallest mean-squared error in estimating the timing of the prosodic peak.

For S to be independent of P, it must meet three conditions:

1. The prosodic peak position must have no intrinsic dependence on the segmental structure. This is true for all superposition models; it would not be true for a model where $p(t)$ explicitly depends on the segments.

2. The corpus must have no built-in correlations between the segmental structure and the underlying prosodic contour (i.e. with the expected type of accent). This should

be approximately true for the corpus used here, especially after we split the corpus into sub-corpora that have uniform frames. See §III.B and §III.D.

3. The algorithm must have no dependence on the segmental structure. Any algorithm Q has that desirable property if (1) it is computed just from f_0, and (2) if the algorithm does not depend on the choice of origin for the time axis, i.e. $Q\{f(t+\eta)\} = Q\{f(t)\}+\eta$ for any η. Thus, if one were to delay the speech by 1 second, the result of the algorithm f_0 should also be delayed by 1 second. All the algorithms we consider have both properties.

Thus, the algorithm that has the smallest variance of its estimates will have the smallest errors between its estimates and the underlying position of the peaks of the prosodic contour. This will be true even if there is some intrinsic variability in the underlying prosody.

APPENDIX B: EDGE CASES

The discussion of the operation of the bracketed maximum algorithm in §II.D assumed voicing everywhere. This section discusses cases where unvoiced regions are important. Ultimately, though, its value will be tested experimentally in §III.

1. Prosodic Maximum within, but near the edge of voiced region

Figure 14 shows sample data annotated to show the operation of the bracketed maximum algorithm. In this situation, one of the bracketing measurements may end up at the voiced/unvoiced transition. Under some common conditions, such as a large positive segmental shift at the edge of a voiced region (e.g. van Santen and Hirschberg, 1994, Figure 4), the bracketed maximum should provide a substantially better estimate of the position of the peak of $p(t)$ than would the simple maximum. However, the average accuracy relative to the simple maximum is not trivial to predict.

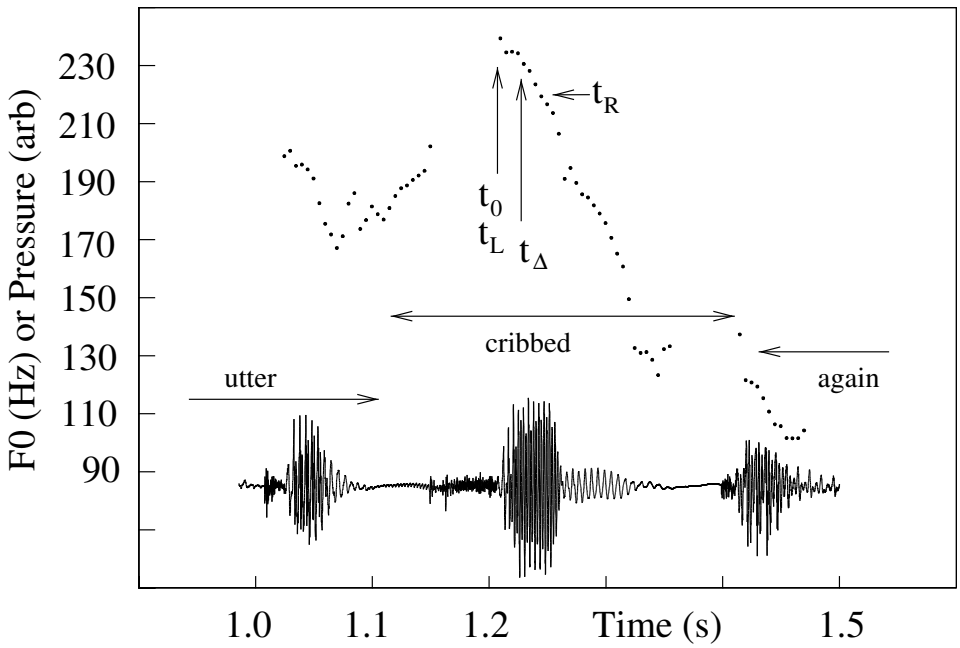

FIG. 14. Sample audio data, plotted as Figure 6. This data was chosen to show the operation of the bracketed maximum algorithm, when the f_0 maximum is at the edge of a voiced region.

2. Prosodic Maximum in unvoiced region

In this case, neither the simple maximum nor the algorithm presented here can accurately mark the prosodic peak. The bracketed maximum will typically be less accurate than the simple maximum, as it is biassed away from unvoiced regions. However, from the point of view of human-to-human communication, this ought not to be an important case. If the precise timing of f_0 peaks is indeed an important part of the language, it seems unlikely that the language's phonological rules would evolve in such a way as to put the peaks where they cannot be heard.

APPENDIX C: SECTIONS OF THE CORPUS

The corpus (§III.B) has a mixture of four different frames, and Figure 13 shows the average over the entire corpus However the different frames could have different behaviours. We check this by re-plotting Figure 13 with the data from each frame separately.

Figures 15 and 16 show that the various components of the corpus have qualitatively similar behaviours, especially for the bracketed maximum algorithm. Variances are generally lower for words beginning with a consonant than for those that begin with vowels. (These words are also more common, so they dominate the overall average.) Possibly, f_0 shifts caused by consonants constrain the peak position, thus reducing the variance. Figure 16 shows somewhat different behaviour for Cx words vs. Vx words: the variance of the latter rises when a large smoothing window is used.

ENDNOTES

1. This term assumes that there is an underlying prosody that you are attempting to produce (e.g. this is your linguistic competence), but that the actual performance is disturbed by other effects. "Prosodic peak" refers to the peak of the underlying prosody. Equation 1 provides a mathematical model of this idea.

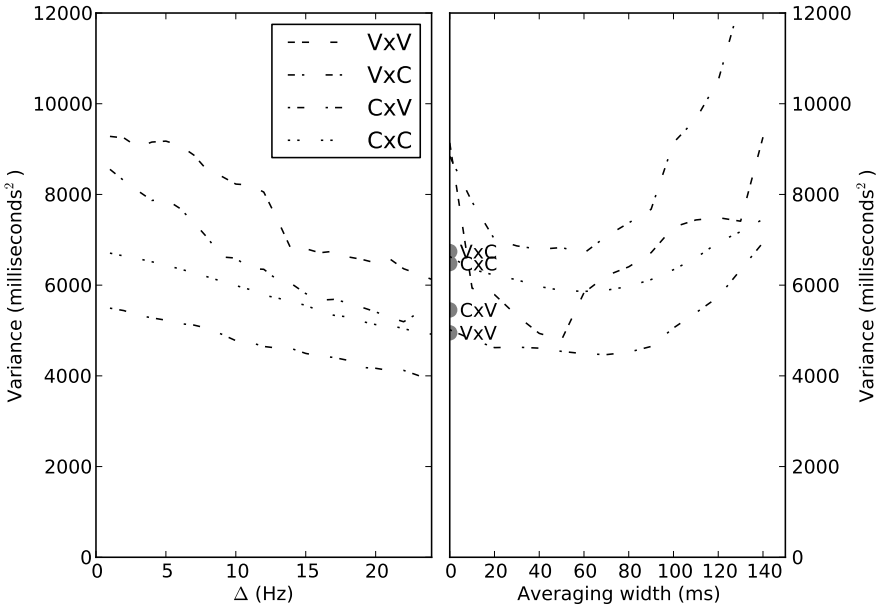

FIG. 15. Left: variance of the bracketed maximum estimate of the prosodic peak position, plotted against Δ. The lines show the variance of each section; the dot-dash pattern indicates the form of the word (§III.B). Right: variance of algorithms that median-smooth the f_0 data and then take the simple maximum. The variance is plotted against the width of the smoothing window. (The appropriately weighted average of these curves corresponds to the dot-dashed line in Figure 13, right.) Results for MOMEL are shown by the large grey dots, and are labeled by the form of the word.

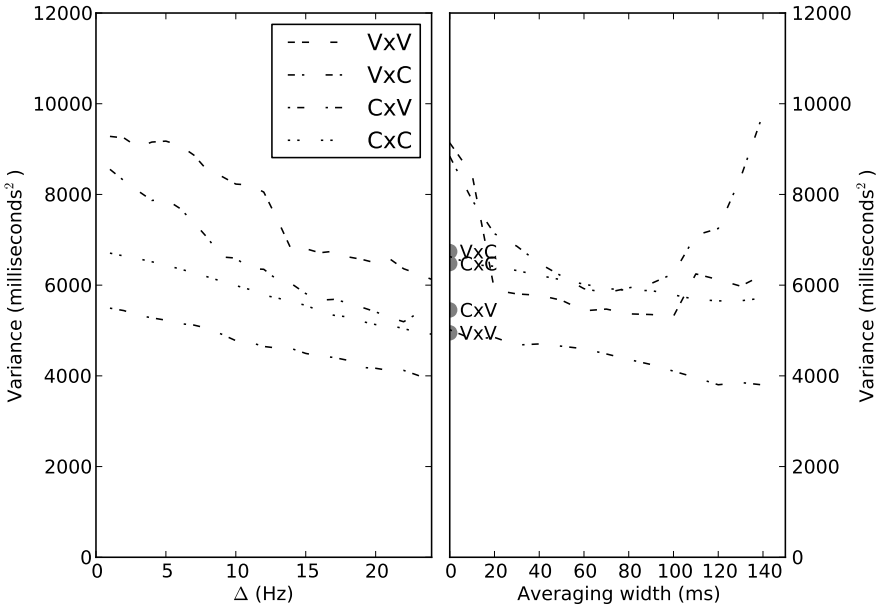

FIG. 16. Left: variance of the bracketed maximum, plotted as per Figure 15. Right: variance of algorithms that mean-smooth the f_0 data and then take the simple maximum.

2. For instance, one cannot describe animals with a superposition model of heads onto bodies. This is because some combinations, such as putting a horse's head on a hummingbird's body, do not yield biologically valid animals.

3. The model in van Santen and Hirschberg (1994) and van Santen *et al.* (1998) is not strictly a superposition model because the segmental effects are not independent of the prosodic contour. The prosodic contour in their model explicitly depends upon the segmental structure.

4. This independence assumption implies that segment-related frequency shifts are the same in high f_0 regions as in regions with low f_0. For the purposes of this paper, it does not much matter whether this is true on a linear or a logarithmic (or other) frequency scale.

5. Alternatively, if segmental frequency shifts are a core part of the language, then some of the available intonation information is being used to help listeners discriminate vowels. This alternative implies that, to be unambiguous to a listener, prosodically meaningful pitch motions would be at least as large as segmentally-related f_0 shifts, which would strengthen the arguments in §II.E.

6. However, one can imagine that the example might correspond to a nonsense phrase like "ma**mi**mami**ma**mi..."

7. The possibility of seeing spurious f_0 maxima due to segmental effects was mentioned in van Santen *et al.* (1998).

8. For example, Gussenhoven (1999) argues (based upon Pierrehumbert and Steele, 1989) that "..., if subjects were to produce a bimodal distribution of peak times in their imitations, then the difference must be categorical." Their deduction can now be seen to be incorrect, given the counterexample presented here that is bimodal but not discrete.

9. The misinterpretation need not be on the part of a linguist. To the extent that peak positions are perceptually important, the listener might also perceive a sharp, phonological distinction that was not necessarily intended by the speaker. Possibly, this effect could lead to listeners later intentionally producing a bimodal distribution of peak positions, which might eventually lead to a phonological distinction becoming part of the language. Such an effect is presumably most likely on common phrases, where the same segmental structure might be used with a variety of intonation patterns. (Private communication, Elinor Payne, 9/2008.)

10. By "simple maximum," I mean the time of the peak f_0. Mathematically, this is often written as $\text{argmax}\{f(t)\}$.

11. Hawkins and White (1988) argues that the the builders of Stonehenge used this idea to mark the day of most northerly sunrise. So, this technique may have been discovered before c. 2800 BCE.

12. Two measurements are approximately independent if knowledge of one thing is only very slightly useful in predicting the other.

13. If declination and/or down-step were well enough understood to be predictable without reference to the speech data, it might make sense to use a larger Δ on the forwards side. However, such an algorithm might misbehave badly at places where the right side of a peak is higher than the left.

14. Note that these experiments involve comparisons of sentences with identical sequences of phones, so the listener does not need to compensate for segmental effects. The listener's task in these experiments is thus easier than in natural speech, and these discrimination thresholds should thus provide a lower bound on how much f_0 shift is necessary to indicate an accent under more realistic conditions.

15. Note that high f_0 peaks are neither the only nor the best indicator of prominence. See Kochanski *et al.* (2005) for discussion of this point.

REFERENCES

Abel, J. S. (**1990**). "Optimal sensor placement for passive source localization", in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90*, volume 5, 2927–2930, Albuquerque, NM, USA, 3-6 April 1990.

Arvaniti, A., Ladd, D. R., and Mennen, I. (**1998**). "Stability of tonal alignment: the case of Greek prenuclear accents", J. Phonetics **26**, 3–25.

Baken, R. J. (**1987**). *Clinical Measurement of Speech and Voice* (Taylor and Francis, London).

Braun, B., Kochanski, G., Grabe, E., and Rosner, B. S. (**2006**). "Evidence for attractors in English intonation", J. Acoustical Society of America **119**, 4006–4015, URL `doi:10.1121/1.2195267`.

Chen, A., Gussenhoven, C., and Rietveld, T. (**2004**). "Language-specificity in the perception of paralinguistic intonational meaning", Language and Speech **47**, 311–349.

Chuang, C.-K. and Wang, W. S.-Y. (**1978**). "Psychophysical pitch biases related to vowel quality, intensity difference, and sequential order", J. Acoustical Society of America **64**, 1004–1014, URL `http://dx.doi.org/10.1121/1.382083`.

Crandall, J. B. (**1925**). "The sounds of speech", Bell System Technical Journal **4**, 586–626.

Di Cristo, A. and Hirst, D. J. (**1986**). "Modeling French micromelody: analysis and synthesis", Phonetica **43**, 11–30.

Dilley, L. C., Ladd, D. R., and Schepman, A. (**2005**). "Alignment of L and H in bitonal pitch accents: testing two hypotheses", J. Phonetics **33**, 115–119.

Fujisaki, H. (**1983**). "Dynamic characteristics of voice fundamental frequency in speech and singing", in *The Production of Speech*, edited by P. F. MacNeilage, 39–55 (Springer, New York).

Greenberg, S., Chang, S., and Hitchcock, L. (**2001**). "The relation between stress accent and vocalic identity in spontaneous American English discourse", in *Prosody-2001*, Paper 9 (International Speech Communication Organization, `http://www.isca-speech.org/`

contact.html), URL http://www.isca-speech.org/archive/prosody_2001/prsr_009.html, viewed 10/2008; presented at the conference "Prosody in Speech Recognition and Understanding", October 22-24, 2001, Red Bank, NJ, USA.

Gussenhoven, C. (**1999**). "Discreteness and gradience in intonational contrasts", Language and Speech **42**, 283–305.

Hawkins, G. S. and White, J. B. (**1988**). *Stonehenge Decoded* (Hippocrene Books, New York).

Hirst, D. and Espesser, R. (**1993**). "Automatic modelling of fundamental frequency using a quadratic spline function", in *Travaux de l'Institute de Phonétique d'Aix*, volume 15, 75–85 (Universite de Provence, Aix de Provence, France), U.R.A. 261 CNRS.

Holm, B. and Bailly, G. (**2002**). "Learning the hidden structure of intonation: Implementing various functions of prosody", in *Proceedings of Speech Prosody 2002*, URL http://www.isca-speech.org/archive/sp2002/sp02_399.pdf.

House, A. S. and Fairbanks, G. (**1953**). "The influence of consonant environment upon the secondary acoustic characteristics of vowels", J. Acoustical Society of America **25**, 105–113.

House, D. (**2003**). "Perceiving question intonation: the role of pre-focal pause and delayed focal peak", in *Proceedings of the 15*$^{\text{th}}$ *International Congress of the Phonetic Sciences (ICPhS)*, 755–758 (Barcelona).

House, J., Dankovicova, J., and Huckvale, M. (**1999**). "An integrated prosodic approach to speech synthesis", in *Proceedings of the XIV ICPhS (International Congress of the Phonetic Sciences)*, volume 3, 2343–2346 (The University of California, Berkeley, California), San Francisco, USA.

Knight, R.-A. (**2002**). "The influence of pitch span on intonational plateaux", in *Proceedings of the Speech Prosody Conference*, 439–442 (Laboratoire Parole et Langage, Université de Provence, France, Aix en Provence).

Knight, R.-A. and Nolan, F. (**2006**). "The effect of pitch span on intonational plateaux", Journal of the International Phonetic Association **36**, 21–38.

Kochanski, G., Grabe, E., Coleman, J., and Rosner, B. (**2005**). "Loudness predicts

prominence: Fundamental frequency lends little", J. Acoustical Society of America **118**, 1038–1054.

Kochanski, G. and Orphanidou, C. (**2008**). "What marks the beat of speech?", J. Acoustical Society of America **123**, 2780–2791, URL `http://kochanski.org/gpk/papers/2006tapping.pdf`, URL viewed 7/2008.

Krolik, J. L. (**1996**). "The performance of matched-field beamformers with mediterranean-vertical array data", IEEE Transactions on Signal Processing **44**, 2605–2611.

Ladd, D. R. (**1996**). *Intonational Phonology* (Cambridge University Press, Cambridge).

Ladd, D. R., Faulkner, D., Faulkner, H., and Schepman, A. (**1999**). "Constant 'segmental anchoring' of f_0 movements under changes in speech rate", J. Acoustical Society of America **106**, 1543–1554.

Ladd, D. R., Mennen, I., and Schepman, A. (**2000**). "Phonological conditioning of peak alignment in rising pitch accents in Dutch", J. Acoustical Society of America **107**, 2685–2696.

Ladefoged, P. (**1964**). "Some possibilities in speech synthesis", Language and Speech **7**, 205–214.

Mack, M. A. and Gold, B. (**1984**). "The discrimination of pitch in pulse trains and speech", Technical Report 680, Lincoln Laboratory, Massachusetts Institute of Technology, URL `http://handle.dtic.mil/100.2/ADA142996`, Defense Technical Information Center accession number AD-A142 996; Prepared for the Department of the Air Force under Electronic Systems Division Contract F19628-80-C-0002. URL viewed 4/2008.

Morlec, Y., Bailly, B., and Aubergé, V. (**1996**). "Generating intonation by superposing gestures", in *Proceedings of the Fourth International Conference on Spoken Language Processing, ICSLP 96* (International Speech Communications Association, http://www.isca-speech.org).

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dancovicova, J., and Heid, S. (**2000**). "Prosynth: an integrated prosodic approach to device independent, natural-sounding speech synthesis", Computer Speech and Language **14**, 177–210.

Peng, S.-H. (**2000**). "Lexical versus 'phonological' representations of Mandarin sandhi

tones", in *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, edited by M. B. Broe and J. B. Pierrehumbert, 152–167 (Cambridge University Press, Cambridge, UK).

Peterson, G. E. and Barney, H. L. (**1952**). "Control methods used in a study of the vowels", J. Acoustical Society of America **24**, 175–184.

Pierrehumbert, J. B. and Steele, S. A. (**1989**). "Categories of tonal alignment in English", Phonetica **46**, 181–196.

Ross, K. N. and Ostendorf, M. (**1999**). "A dynamical system model for generating fundamental frequency for speech synthesis", IEEE Transactions on Speech and Audio Processing **7**, 295–309.

Sachs, T. S., Meyer, C. H., Irarrazabal, P., Hu, B. S., Nishimura, D. G., and Macovski, A. (**1995**). "The diminishing variance algorithm for real-time reduction of motion artifacts in MRI", Magnetic Resonance in Medicine 412–422.

Silverman, K. (**1987**). "The structure and processing of fundamental frequency contours", Ph.D. thesis, Cambridge University, Cambridge, UK.

Silverman, K. and Pierrehumbert, J. (**1990**). "The timing of prenuclear high accents in English", in *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, edited by J. Kingston and M. Beckman, 72–106 (Cambridge University Press, Cambridge).

Slater, A. and Coleman, J. (**1996**). "Non-segmental analysis and synthesis based on a speech database", in *Proceedings of ICSLP 96, Fourth International Conference on Spoken Language Processing*, edited by H. T. Bunnell and W. Idsardi, volume 4, 2379–2382.

Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT press, Cambridge, MA).

Taylor, H. C. (**1933**). "The fundamental pitch of English vowels", J. Experimental Psychology **16**, 565–582.

Taylor, P. (**1993**). "Automatic recognition of intonation from f_0 contours using the rise/fall/connection model", in *Proceedings of the Third European Conference on Speech Communication and Technology (EUROSPEECH '93)*, volume 2 (International Speech Communications Organization, `http://www.isca-speech.org`), URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.2522`, viewed 10/2008.

van Santen, J. and Hirschberg, J. (**1994**). "Segmental effects on timing and height of pitch contours", in *Proceedings of ICSLP-94, Third International Conference on Spoken Language Processing*, 719–722, Yokohama.

van Santen, J. P. H. and Möbius, B. (**1999**). "A quantitative model of f_0 generation and alignment", in *Intonation Analysis, Modelling and Technology*, edited by A. Botinis, 269–288 (Kluwer Academic Publishers, Dordrecht, Netherlands).

van Santen, J. P. H., Möbius, B., Venditti, J. J., and Shih, C. (**1998**). "Description of the Bell Labs intonation system", in *Proceedings of the Third International Workshop on Speech Synthesis*, 293–398 (International Speech Communications Organization, `http://www.isca-speech.org`), URL `http://www.slt.atr.co.jp/cocosda/jenolan/Proc/r84/r84.pdf`, jenolan Caves, Australia; URL viewed 10/2008.

Whalen, D. H. and Levitt, A. G. (**1995**). "The universality of intrinsic f_0 of vowels", J. of Phonetics **17**, 193–203.

Wichmann, A., House, J., and Rietveld, T. (**1997**). "Peak displacement and topic structure", in *Intonation: Theory, Models, and Applications* (International Speech Communications Organization, `http://www.isca-speech.org`), URL `http://isca-speech.org/archive/int_97/inta_329.html`.

Xu, C. X. and Xu, Y. (**2003**). "f_0 perturbations by consonants and their implications on tone recognition", in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 456–459 (IEEE, `http://ieee.org`), ICASSP '03.

Xu, Y. and Sun, X. J. (**2000**). "How fast can we really change pitch? Maximum speed of pitch change revisited.", in *Proceedings of the International Conference on Spoken Language Processing 2000*.

LIST OF FIGURES

FIG. 1 The underlying intonation contour $p(t)$ (dashed) and the surface intonation $f(t)$ (solid line) for the model utterance. The plots shows fundamental frequency vs. time. The difference between the contours is the segmental effect, $s(t)$. 7

FIG. 2 A set of contours for the intonation component, $p(t)$. These form inputs to the intonation model, Equation 1. The contours are shifted vertically for clarity. Contours are labelled by τ (in seconds) and the maxima of neighboring curves differ by 30 ms. 8

FIG. 3 A set of contours for $f(t)$, derived from Figure 2 by adding segmental effects (Equation 2). Small circles mark the maxima of the contours. Plotted as per Figure 2. 9

FIG. 4 The relationship between the time shift τ of the underlying intonation contour and the maximum of the observed fundamental frequency, argmax$\{f(t)\}$. The vertical axis indicates the position of the maxima (open circles) in Figure 3. The diagonal dashed line corresponds to the position of the underlying prosodic peak. 11

FIG. 5 The distribution of the times of f_0 peaks including segmental shifts (dots). This histogram assumes a Gaussian distribution of τ (the location of the underlying intonation maxima) with a standard deviation of 40 milliseconds. (A shifted version of the distribution of τ is shown by the thin dashed line.) The horizontal axis is the position of the maximum and the vertical axis is the probability of observation, collected into 10 millisecond bins. 14

FIG. 6 Sample audio data is plotted (lower curve) with a transcription, and an f_0 contour above that. The f_0 contour is annotated with the simple maximum (labeled t_0), intermediate results from the bracketed maximum algorithm with $\Delta = 20$ Hz (labeled t_L and t_R), and the final result (labeled t_Δ). This data was chosen to show the operation of the bracketed maximum algorithm on a broad f_0 peak. 17

FIG. 7 The bracketed maximum (solid), the simple maximum (dashed) *vs.* τ, the maximum of the underlying intonation contour. 18

FIG. 8 The distribution of estimated intonation peak positions, $(t_L + t_R)/2$, for a simulated experiment using the bracketed maximum technique (solid line w/dots). The resulting histogram is unimodal, reflecting the unimodal distribution of τ (alignment). Results from the simple maximum (from Figure 5) are reproduced as a dotted line for comparison, and the underlying distribution of the prosodic peak position (τ) is shown as a dashed line. . . 19

FIG. 9 Quasi-duration scatter-plot for the entire corpus. The quasi-duration is a measure of the stationarity of the speech spectrum; small values imply a rapidly changing spectrum. The plot shows the log of the normalized quasi-duration against normalized time, otherwise plotted as per Figure 11. . . . 24

FIG. 10 Normalized loudness vs. normalized time for the entire corpus. Plotted as Figure 11, except that all dots are the same size. 25

FIG. 11 Fundamental frequency of the entire corpus. This plots normalized f_0 against normalized time for the region of interest. The horizontal axis goes from the beginning of the syllable before the variable syllable to the end of the syllable after. The vertical axis is f_0 deviation from 170 Hz, in semitones. This is a smoothed scatterplot of f_0 measurements. (The vertical stripes, e.g. near $x = 0.5$ are the result of the 10 ms interval between f_0 measurements.) . . 26

FIG. 12 Sample flat-topped f_0 profile. 27

FIG. 13 Left: variance of the bracketed maximum estimate of the prosodic peak position, plotted against Δ. The solid line shows the variance of the entire corpus. The simple maximum is shown by the dash-dot line, for comparison. Right: variance of algorithms that smooth the f_0 data and then take the maximum. The variance is plotted against the width of the smoothing window. The best MOMEL result is the grey dot. (MOMEL does not have a comparable smoothing parameter, so the horizontal axis is irrelevant.) . . 29

FIG. 14 Sample audio data, plotted as Figure 6. This data was chosen to show the operation of the bracketed maximum algorithm, when the f_0 maximum is at the edge of a voiced region. 36

FIG. 15 Left: variance of the bracketed maximum estimate of the prosodic peak position, plotted against Δ. The lines show the variance of each section; the dot-dash pattern indicates the form of the word (§III.B). Right: variance of algorithms that median-smooth the f_0 data and then take the simple maximum. The variance is plotted against the width of the smoothing window. (The appropriately weighted average of these curves corresponds to the dot-dashed line in Figure 13, right.) Results for MOMEL are shown by the large grey dots, and are labeled by the form of the word. 38

FIG. 16 Left: variance of the bracketed maximum, plotted as per Figure 15. Right: variance of algorithms that mean-smooth the f_0 data and then take the simple maximum. 39