“It certainly seems unlikely that $F_0$ turning points are directly perceived as such. … It … seems more probable that, in some way, the location of the turning point affects the perception of the $F_0$ level of the accented syllables. These questions are only just beginning to be investigated empirically.” (Ladd 2008 [1996]: 180)

**Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology**

Abstract

Recent evidence that pitch-movement shape can influence perceived alignment of rising (LH) pitch accents in several languages appears to challenge the well-established level-based approach to intonation embodied in the AM model, wherein it is typically assumed that the alignment and scaling of well-defined turning points (TPs) in the $F_0$ contour are the primary phonetic correlates of contrastive accent category. Here we present the results of 2 experiments, arguing that a new approach to tonal implementation succeeds in reconciling these apparent contradictions. This approach, based on the notion of a perceptual reference point called Tonal Center of Gravity (TCoG), treats information about contour shape and TP-localization not in ‘either-or’ terms, but rather as two sets of cues working in a fundamentally synergistic way toward a single perceptual end: the alignment and scaling of TCoG. Experiment 1 shows that TCoG-based models can perform better at distinguishing productions of English L+H* and L*+H pitch accents than comparable TP-only-based models; Experiment 2 shows that TCoG is more robust than TP-only-based models to ambiguities in TP localization commonly encountered in $F_0$ signals from natural speech. TCoG is shown to capture key insights of movement-based approaches to intonation, without abandoning the central advantages of level-based approaches like AM.
I. Introduction

At least since Bolinger (1951), it has been customary to view intonational theory in terms of the competition between two opposing views of the nature and implementation of the primitive elements from which intonation contours are constructed: Configurationist models, based on explicitly defined $F_0$ movement shapes (e.g., rises, falls, ‘sustentions’) and level-based models, which view intonation contours as sequences of tone-level targets, such as Highs and Lows (or levels /1/, /2/, /3/, /4/, etc.). Perhaps the most influential linguistic theory of intonation today, the Autosegmental-Metrical (AM) model (Pierrehumbert 1980; Ladd 2008), is of the level-based type: target intonation contours are seen as strings of (mostly) High and Low phonological tone-level specifications, associated with particular positions or hosts in the segmental string. Between these specified locations lie tonally unspecified stretches that can sometimes be quite lengthy. Phonetically, tone specifications are realized by highly localized, systematically positioned $F_0$ targets, between which the $F_0$ curve is held to be simply an interpolation along something like the simplest or articulatorily cheapest path. This paper explores a new approach to $F_0$ targets and their implementation, one that is compatible with a level-based phonology such as that of AM, while at the same time accounting for the demonstrated relevance of global contour shape in tonal perception and production.

1.1. Turning points and tonal targets: Phonetic support for a level-based phonology

In the context of the standard target-and-interpolation model, the phonetic realization of intonational pitch events is typically investigated in terms of the localization of observable turning points in the $F_0$ curve (e.g., maxima, minima, ‘elbows’; hereafter TPs), along two critical dimensions: 1) the timing, or alignment, of tonal targets with respect to the segmental string (e.g., Arvaniti, Ladd, and Mennen 1998; Ladd et al. 1999; Ladd, Mennen, and Schepman 2000; Ladd and Schepman 2003, *inter alia*), and 2) the realization, or scaling, of tonal targets in the ‘vertical’ or frequency domain (e.g., Liberman and Pierrehumbert 1984; Ladd 1988, 1990; among many others). Most researchers are careful to avoid positing ‘identity’ between phonological tones and the TPs that are seen as realizing them.\(^1\) However, the ready accessibility of TPs as phonetic reflections of underlying tonal representations has led to the systematic application, within AM description and argumentation, of a kind of analytic inference from a stably observable ‘corner’ in an $F_0$ trace to a tonal target in the phonology posited to account for it. What might otherwise seem purely a matter of technical convenience has therefore become a matter of consequence for phonology as well.

The investigations of alignment and scaling properties of $F_0$ TPs that have resulted from this approach have yielded important advances in our understanding of tonal implementation. Perhaps chief among
these is the strong support these results lend to the case for a level-based approach to intonational phonology. The key finding here is what has come to be known as the *segmental anchoring* of F0 TPs (Arvaniti, Ladd, and Mennen 1998, and many since). Segmental anchoring refers to the state of affairs whereby, under changes to the durations or compositions of regions of the segmental string spanned by intonational F0 movements, the alignment of the F0 TPs delimiting those movements stays relatively stable with respect to select segmental landmarks in the area. In contrast, the shape of the pitch movement realizing a given string of tonal targets (usually construed in terms of its duration or slope) is seen to be highly variable. This pattern is depicted schematically in Figure 1.

![Figure 1](image.png)

*Figure 1.* A schematic depicting the relative stability of the alignment of F0 movements (blue line) with respect to the segmental string (here CVC), and the accompanying variation in the shape (i.e. slope and duration) of the F0 movement.

The apparent lack of invariance in the shapes of intonational pitch movements that emerges here seems at odds with the predictions of configuration-based approaches, while according nicely with the insight lying at the heart of the AM tradition, famously presaged by Bruce (1977), who observed that, for contrasting accent categories in Swedish, “reaching a certain pitch level at a particular point in time is the important thing, not the movement (rise or fall) itself”. Ladd (2008: 67) describes the essence of Bruce’s discovery thus: “… the F0 configurations that happen to span the accented syllables play no useful role in phonetic description of the overall contour; the invariant features of the pitch system appear to be the *turning points* in the contour rather than the transitions that connect them."

1.2. Can TPs tell the whole story?

Alongside this success, however, come serious questions about the role of TPs in the theory of tonal implementation. The first of these problems deals with the recoverability of TPs as cues to tonal representations. Though some TPs are plainly visible in the F0 record, in other cases the precise locations of F0 TPs can be extremely ambiguous, so that no single point in the contour stands out from the others in such a way as to be straightforwardly identified as the ‘target’ in question. This problem is well-known from the case of high accentual plateaux (D’Imperio 2000; Knight 2008), and arises particularly frequently in the case of Low tone specifications as well. Often, for example, instead of an easily isolable
minimum flanked by sharp falling and rising regions, what we find are extended low regions, bounded by shallow rises or falls, such that neither a unique minimum, nor any clear inflection point demarcating movement offsets or onsets, can be identified. (An illustrative example is the F0 ‘elbows’ sometimes identified with L- phrase accents, on which see Barnes et al., 2010a). The fact that analytic procedures have been developed for dealing with such cases in the context of prosodic labelling does nothing to lessen the mystery surrounding how listeners cope with such cases during perception.

Worse still, regions that could define crucial TPs are also frequently simply missing from the F0 curve owing to, e.g., voicelessness or irregular phonation. It might be imagined that human listeners deal with this problem by somehow extrapolating available F0 trajectories in order to perceptually ‘restore’ missing TPs such as peaks or valleys.2 Psychoacoustic inquiries into the matter, however, suggest that this is not the case. Evidence against the idea of extrapolation based on observed F0 trajectories has been available for some time (Dannenbring 1976; Ciocca and Bregman 1987; Bregman 1994). More recently, Barnes and colleagues (2011a) provide evidence that, at least as concerns the perception of pitch-accent scaling, no perceptual completion of any kind appears to take place. Instead of extrapolating or interpolating across voiceless intervals, listeners appear simply to ignore these gaps, making judgments about scaling solely on the basis of F0 information actually present in the signal. To the extent that interruptions in the F0 track seem to cause no reduction in the intelligibility of the speech in question, this data presents a challenge for theories of intonational perception based on the localization of TPs that ought otherwise to have fallen within the missing F0 region.

The biggest problem for strictly TP-based approaches, however, is an increasing body of evidence demonstrating that a range of difficult-to-quantify aspects of global contour shape play a powerful role in the perception of intonational contrasts, often apparently overriding evidence from TPs in determining listener judgments of category membership. These include: ‘peak’ shape (i.e. sharp peak vs. plateau: ’t Hart 1991; D’Imperio 2000: 167-170; Knight 2008); rise or fall duration (in particular, cases in which the TPs in question do not correspond to any of the tonal targets normally thought to ‘constitute’ the F0 events in question, e.g., the end of the fall following an L*+H pitch accent, or the beginning of the rise preceding an H*+L: D’Imperio 2000: 175-176; Niebuhr 2007a); pitch movement curvature (e.g., Welby 2003; Dombrowski and Niebuhr 2005; Barnes, et al., 2010b); and the relative frequency scaling of pitch movement onsets and offsets (D’Imperio 2000: 178). We review each of these problems in more detail in Section 3. For now, by way of clarifying the nature of the challenge, we will limit ourselves to consideration of just the last of these observations.

The facts of the case are as follows: D’Imperio (2000) conducted a series of perceptual studies in which Neapolitan Italian listeners were required to categorize synthetic pitch accents with differing shape and alignment properties as either earlier- or later-timed with respect to the pitch-accented syllable (i.e.
L+H* vs. L*+H in ToBI terms). What she found (among other things) was that, in comparison with a symmetrical rise-fall baseline accent shape (the dashed line in Figure 2), an asymmetrically scaled rise-fall shape (the solid line in the left display) systematically biased listeners toward judgments of later tonal timing (that is, toward L*+H). The mirror image of this manipulation (i.e. with the rise beginning higher than the end of the fall, the solid line in the right display) biased listeners in the opposite direction (toward earlier timing judgments, or L+H*).

![Figure 2](image)

*Figure 2.* A schematic showing how asymmetrically scaled rise-fall shapes systematically biased listeners judgments of alignment (after D’Imperio 2000). Perceived alignment moved in the direction of the arrow when listeners heard an F0 contour like the one schematized by the dark continuous line, compared to one corresponding to the dotted line.

This result is troubling from the point of view of a strictly TP-based model for a number of reasons. Most obviously, though the contrast between L+H* and L*+H is typically construed as one of differing TP timing patterns (and indeed, differences in TP timing alone are sufficient to cue it), the manipulations depicted above do not involve changes to the alignment of TPs at all. Rather, the differences here are exclusively in the domain of scaling, and stranger still, in the relative scaling of non-adjacent TPs (i.e. the beginning of the rise and the end of the fall). Finally, from the point of view of a TP-based implementation of AM theory, it is unclear why the TP created at the end of the fall should influence the perceived timing of the LH-shaped pitch accent in the first place (to the extent that it reflects a Low tonal target that is not ‘part of’ the pitch accent in the usual sense.)

These results, together with the others referenced above, might seem to raise serious questions for the level-based approach to intonational phonology. Listeners, it appears, are paying attention in perception to precisely those aspects of intonation contours that the target-and-interpolation approach predicts they should ignore. Our first major question, therefore, is how we can reconcile these results with what otherwise appears to be strong evidence in favor of the level-based approach to intonational phonology, since, contrary to the expectations of a theory based solely on straight-line interpolation between TP targets, it is by now clear that contour shape in fact does matter. An equally pressing question is why contour shape should matter in precisely the way that it does. Put another way, what is missing is an explanatory theory of the contribution of contour shape to the perception of intonational contrasts that would allow us to predict which subset, of a seemingly limitless array of potential variations in contour
shape, should influence listeners’ perception, and in what direction that influence should operate.

In this paper we present a new approach to the phonetics and phonology of intonation, designed to overcome the challenges reviewed in the foregoing. This model is based on the notion of Tonal Center of Gravity (TCoG), a gestalt or global measure of F₀ event localization that succeeds in accounting both for the demonstrated contributions of F₀ TP-alignment, and for the strength of global F₀ contour shape as cues to intonational contrasts, while referring directly to neither of these things. The TCoG model lies at the heart of a research program whose goal, broadly expressed, is to develop a more robust and perceptually realistic model of tonal timing and scaling patterns than currently exists; one that captures key configurationist insights (i.e. the relevance of contour shape in tonal implementation), but nonetheless maintains the core advantages of a level-based AM phonology.

1.3. An alternative to F₀ turning points: Introducing Tonal Center of Gravity

As noted, the TCoG model is designed in part to reconcile apparently contradictory findings from the experimental literature on intonational primitives. In particular, the lack of shape invariance demonstrated in the segmental anchoring literature sits uneasily beside the demonstrated perceptual relevance of aspects of global contour shape in the perception of intonation. Investigating contour shape systematically, however, is a complex undertaking. One immediate problem we encounter is the apparent heterogeneity of the aspects of contour shape that have been shown to bias, for example, the perception of tonal timing, in one direction or another. As reviewed briefly above, various aspects of the durations, curvatures, slopes, and symmetries of the rises and falls comprising F₀ peaks have been implicated in the perception of H tone timing, but how to generalize across these findings?

The fundamental insight TCoG brings to these problems is the following: Instead of understanding the timing and scaling of F₀ events in terms of the precise locations of any particular geometrical point or points within the F₀ contour, TCoG represents a generalization about the overall distribution of the ‘weight’ or ‘mass’ of the F₀ event in question. In the context of the timing of a High pitch accent, then, the critical issue is the disposition of the bulk of the High F₀ region with respect to the pitch-accented syllable. This idea can be operationalized using a weighted average, as given in Equation 1. Here, the model computes an average of discrete time values at sample locations within a given region of interest, weighted with their measured F₀. The result is a single time value corresponding to the ‘center of gravity’ of the F₀ event in question, a point that can serve as a reference location for that F₀ event in perception.
Equation 1. Weighted average deriving the location of TCoG in time.

\[ TCoG = \frac{\sum_i F0_i t_i}{\sum_i F0_i} \]

The key point in understanding how temporal TCoG captures tonal timing contrasts is that, owing to the nature of the weighted average, intervals on the curve that have high F_0 values contribute more to the outcome than do low F_0 intervals; TCoG is thus drawn toward these high F_0 regions.\(^7\) To see how this works, consider again the asymmetrical rise-fall patterns from D’Imperio’s Neapolitan study, reproduced below in Figure 3 with the location of TCoG in time added for each contour shape. First, in the case of the symmetrical rise-fall, to the extent that the bulk of raised F_0 is distributed evenly to either side of the F_0 maximum within the region of interest, TCoG will in this case coincide perfectly with the accentual peak. If however, we disrupt that symmetry in any way, the balance point will shift: Where the offset of the fall is higher than the onset of the rise (which we will call the ‘right-weighted’ pattern; Figure 3, left panel), the area under the F_0 curve is now greater after the peak than before it, causing TCoG to shift later. Note that a later TCoG predicts the percept of later tonal timing, even in the absence of changes to the temporal alignment of F_0 TPs. This is precisely what D’Imperio documents for her Neapolitan listeners. Likewise, in the mirror image manipulation shown in the right panel of Figure 3, the onset of the rise is higher than the offset of the fall (a ‘left-weighted’ pattern); the bulk of the high F_0 shifts to yield an earlier TCoG, and hence a bias in perception toward judgments of earlier tonal timing. Again, this is just what D’Imperio demonstrates.\(^8\)

\[ \text{Figure 3. A schematic showing how asymmetrically scaled rise-fall shapes systematically bias listeners’ judgments of alignment (after D’Imperio 2000 and Figure 1 above), and how the consequent changes in location of TCoG parallel those perceptual judgments. Dotted lines represent F_0 tracks and TCoG locations for the symmetrical reference examples; unbroken lines represent F_0 tracks and TCoG locations for the asymmetrical ‘right-weighted’ (left panel) and left-weighted (right panel) examples.} \]

Of course, the same shifts in TCoG location could have been accomplished by other means as well. In Section 3 below, we review in detail how TCoG correctly predicts the complete set of global contour shape effects on the perception of tonal timing listed above. For now, we note only that what makes such
a unified account of these phenomena possible is a key difference between the way TCoG approaches contour shape, and what a standard configurationist approach assumes. Specifically, where the core of the configurationist enterprise is the idea that contour shapes themselves are somehow specified or invariant, TCoG explicitly predicts the functional (or better, linguistic) equivalence of contours that have very different shapes, providing the location of their TCoG is the same in what we might call two-dimensional tono-temporal space. We believe this fundamental prediction is correct, i.e. that ‘the same’ contour is indeed often realized by F_0 curves of sometimes-strikingly-differing shapes and TP-alignments that share TCoG location.\(^9\)

Put another way, one means for achieving a later TCoG for a given pitch accent would be to right-weight a rise-fall shape using the relative scaling of the rise onset and fall offset, as depicted above. The same effect, however, might equally be achieved by altering the shape of the rise to be more concave upward (i.e. ‘scooper’), or by extending the duration of the fall, or by lingering at the peak to form a plateau, or indeed, of course, simply through the manipulation of TP-alignment (i.e. by leaving the F_0 contour shape the same, but starting and ending the relevant pitch movements later). Thus, rather than viewing TP-alignment and contour shape as ‘either-or’ alternatives in the construction of a theory of intonational phonetics and phonology, the TCoG approach sees these two seemingly orthogonal dimensions of F_0 contour production as related in a fundamentally synergistic way. Shape and alignment work together in a mutually enhancing fashion toward a single perceptual end: the alignment and scaling of TCoG.\(^10\)

It should be clear from the above that the TCoG model in no way denies the systematicity of TP-alignment in speech production. On the contrary, we view the results in the segmental anchoring literature demonstrating a lack of contour shape invariance to be essentially decisive in ruling out shape per se as the defining cue to intonational category. It should likewise be clear that TCoG does not entail the introduction of contour shape itself into phonological representations in any direct fashion. Indeed, the most straightforward way of situating TCoG within a larger model of intonational phonetics and phonology would be to conceive of it as a sort of phonetic ‘front end’ for a level-based model like AM. Under this scenario, each tonal specification in an AM string would be realized phonetically in terms of timing and scaling as an F_0 event with a TCoG located at a set of target coordinates within a two-dimensional tono-temporal space. Speakers, however, would be understood to have at their disposal multiple strategies for achieving a given alignment and scaling for each tone’s TCoG, including variation in F_0 TP-alignment, variation in contour shape, and most likely variation in both of these simultaneously. (For recent evidence of precisely the kinds of trading relations between timing and shape patterns that TCoG would predict in speech production, see Niebuhr, et al. 2011.)
To summarize the foregoing, we have argued that a model of tonal implementation relying solely on the realizations of $F_0$ TPs has problems characterizing the mapping between the phonological structures of intonation systems and the phonetics of $F_0$. This is so both because reliable extraction of TP locations in $F_0$ space and time appears for many reasons perceptually implausible, as well as because numerous aspects of $F_0$ contour shape not reducible to TP location apparently nevertheless influence listener categorizations of $F_0$ events in a powerful fashion. The TCoG model presented above allows us to overcome these difficulties.

In what follows, we will argue not only that precisely located $F_0$ TPs are insufficient to characterize the intonational phonetics-phonology mapping, but that they are unnecessary as well. It has previously been demonstrated in a perception study (Barnes, et al. 2010b) that a TCoG-based model can account for the impact of contour shape on listener categorizations of American English pitch accents differing in their characteristic tonal timing patterns ($L^+H^*$ and $L^*+H$) in a way that strictly-TP-based models cannot. In the present study, we aim to demonstrate that TCoG does a better job accounting for details of the phonetic implementation of that contrast in speakers’ productions as well.

2. Two experiments testing the TCoG approach

In this section we present the results of two experiments designed to test the hypothesis that alignment of the Tonal Center of Gravity with the segmental skeleton of a spoken utterance provides a more reliable cue for intonational contrasts than does the alignment of $F_0$ turning points alone. The first experiment (Section 2.1) demonstrates the success of the TCoG model in categorizing two American English pitch accents that contrast in (among other things) the timing of their constituent tonal targets. The second (Section 2.2) compares the response of the TCoG model and an analogous TP-based model to the systematic introduction of instability into the process of determining alignments for crucial points in the $F_0$ curve in each model. The goal of this second experiment is to demonstrate the robustness of the TCoG measure in the face of imperfect or ambiguous $F_0$ tracks, owing to its lack of reliance on precisely located points in the $F_0$ record, such as pitch movement onsets and offsets. Such robustness is critical for models of intonation processing, given the widespread ambiguity of TP locations in continuous speech.

2.1. Experiment 1: Categorization of contrasting pitch accents: L$^+$H$^*$ vs. L$^*$+H

Our first experiment compares the ability of a TCoG-based model and analogous TP-based models to correctly classify utterances of American English according to which of two pitch accents, ToBI’s L$^+$H$^*$ and L$^*$+H, they contain. The L$^+$H$^*$ accent is usually considered to be something like a neutral alignment
of a High tone with a pitch-accented syllable, while L*+H is often thought of as a late-aligned High, also known as a ‘delayed peak’ or ‘scooped rise’ (Pierrehumbert 1980; Ladd 1983; Pierrehumbert and Steele, 1989). These pitch accents are known to differ systematically not only in terms of where their $F_0$ maxima occur with respect to the pitch-accented syllable, but also in the alignment of the onset of the accentual rise: For L*+H pitch accents, at least in tokens where the location of turning points is clear, it can be seen that both the beginning and end of the rise occur systematically later than do analogous TPs in instances of L+H*.12 (See, e.g., Brugos, et al. 2008.) This contrast is exemplified in the context of a generally rise-fall-rise shaped contour below in Figure 4.

![Figure 4](image)

Figure 4. $F_0$ contours corresponding to L+H* (left) and L*+H (right), realized in a phrase with L-H%, in an utterance of the text *there’s a mellower one*. The vowel of the syllable bearing the pitch accent in each contour is designated by the vertical shaded area; note that the $F_0$ rise begins during the accented vowel for L+H* but later for L*+H.

Of course, if the rise of an L*+H pitch accent both begins and ends later than does that of an L+H*, then L*+H’s TCoG will align comparatively later as well, suggesting that a model like TCoG should be well suited to the discrimination of these two accents. Additionally, if speakers also use other, non-TP-based aspects of global contour shape to enhance the alignment cues for TCoG, as posited above, we expect TCoG-based models to perform better still.

To provide a testing ground for the categorization capacities of the TCoG model in the context of tonal timing contrasts like that between English L+H* and L*+H, we elicited a substantial corpus of naturalistic productions of both pitch accents, realized on identical strings, from six native speakers of American English. For all the resulting data, we calculated the location of TCoG for the pitch accent’s High tone using several different algorithms, and conducted an analysis focused on the alignment of TCoG with respect to the accented vowel for each of the contrasting pitch accents. Success in pitch accent categorization was assessed using a logistic regression analysis, as described below. Stimuli were designed to be voiced throughout, and to have minimal segmental effects on $F_0$, so that similar analyses of utterances of the same words could be conducted using the alignments of relevant $F_0$ TPs with respect to segmental anchors, providing a comparison of the relative levels of success achieved by each type of model when tested analogously.
2.1.1. Contours

For elicitation purposes, the two pitch accents under investigation were embedded in otherwise phonologically identical target intonation contours that could plausibly be produced on each selected phrase. The contours chosen were the two well-known rise-fall-rise contours called Incredulity and Uncertainty by Ward and Hirschberg (1985, *et seq.*). The Incredulity contour typically expresses something like amazement or disbelief on the part of the speaker. Uncertainty is subtler, characterized by Ward and Hirschberg as expressing uncertainty on the part of the speaker as to the appropriateness of his or her contribution to the discourse. It is often used in the expression of tentative suggestions, or the offering of possible alternatives. (See 2. below for examples.) Ward and Hirschberg describe these two contours as employing the same pitch accent, L*+H, differing primarily just in pitch range, with Incredulity involving rises of greater magnitude. Many speakers, however, appear to use L*+H pitch accents only in the Uncertainty context, and to select L+H* for Incredulity instead, as in Figure 4 above. Brugos and colleagues (2008) document these pitch accent usage preferences, with both alignment and pitch range differences between the two contours, for three naïve speakers. In ToBI terms, then, the contours of interest in this study were Incredulity realized as L+H* L-H%, and Uncertainty realized as L*+H L-H%. To create natural contexts for the utterance of these two contours, we constructed two sets of short dialogues, one for Incredulity and one for Uncertainty, to be read aloud by an experimenter and each participant. Both contextualizing dialogues involve discussion of a crossword puzzle. For the Incredulity contour, Speaker A, the experimenter, is recounting a story about doing a crossword puzzle with a friend, Bob, who, as can be seen in the examples below, routinely provides shockingly inappropriate responses to requests for help with crossword clues. These responses prompt Speaker B, the participant, to echo Bob’s response incredulously. For the Uncertainty contour, by contrast, Speakers A and B are solving a crossword together. A asks B for help with a clue, and B responds by suggesting a solution, conveying some uncertainty about its appropriateness. Sample dialogues for each context are given below, where the target phrase for the rise-fall-rise contours is in each case *There’s ‘lemonier’*.

1. Incredulity and Uncertainty in action

a) Incredulity: Speaker A: *When I asked for a 10-letter word for ‘sweeter’, Bob said “There’s ‘lemonier’”.*
   Speaker B: *‘There’s ‘lemonier’’!? That’s absurd!*

b) Uncertainty: Speaker A: *Can you think of an 8-letter word for ‘more citrusy’?*
   Speaker B: *There’s ‘lemonier’. Would that work?*
2.1.2. Target items

As illustrated in these examples, target phrases for the contrasting contours were always of the shape *There’s ‘X’*, where X was a one- or two-word phrase with primary stress on the first syllable. Care was taken to ensure that the nuclear pitch accent of the contour coincided with this stressed initial syllable. For single words, such as *lémonier*, target items were chosen such that lexical stress fell unambiguously on this syllable, and thus attracted the nuclear pitch accent when located on these items. Two-word phrases were noun-noun compounds that in English would naturally occur with primary stress on the first syllable of the first member of the compound, e.g., *nàil enàmel*, or *mèlon alèrt*. In order to ensure that participants did indeed place the nuclear pitch accent on the first member of these compounds, two-word targets were all placed in dialogue contexts that would suggest narrow focus on the first member of the compound. The following examples illustrate this strategy. Note capitalization for the same reason.

2. Sample stimuli

a) Incredulity:  
Speaker A: *When I asked for a 6-letter word to go before the word ‘enamel’, he said “There’s ‘NAIL enamel”’.  
Speaker B: “There’s ‘NAIL enamel’”?! Is he crazy?

b) Uncertainty:  
Speaker A: *I need a 5-letter word to go before the word ‘alert’---at least I think it’s ‘alert’*.  
Speaker B: *There’s ‘MELON alert’…is that ok?*

For reasons connected to an unrelated study, target items were also selected to represent a variety of distinct metrical and prosodic structures, where elements that varied included the number of syllables following the nuclear pitch accent (1, *lémon*, 2, *lémony*, and 3, *lémonier*), presence/location of a lexical stress following the nuclear pitch accent (*lémonier* vs. *láminàted* vs. *láminalize*) and presence/location of a word boundary following same (*láminàted vs. líme aròma*, *láminalize vs. lémon alàrm*). There were seven such classes, each represented by four or five phrases, for a total of 29 target phrases. The details of this variation play no specific role in our analysis here, but it is an important feature of the data set nonetheless that by varying parameters such as these, we were able to obtain renditions of our contours that differed significantly in the shapes of the $F_0$ curves that realized them. Would-be models of listener categorization of these contours are thus subjected to a challenging degree of variability in the phonetic realization of each contour. A complete list of target phrases is provided in the Appendix.
2.1.3. Participants

Participants were six native speakers of American English between the ages of 18 and 50, one male and five female. One of the female participants, and the sole male participant, were also among the authors of this study; their data were included only after it was determined that they did not differ from the other participants in the study. All participants gave informed consent for the study, which was approved by the Institutional Review Boards of Boston University, Massachusetts Institute of Technology, and Simmons College.

2.1.4. Elicitation and evaluation of production data

Elicitations were conducted in two separate sessions approximately one week apart. One session was devoted to the Incredulity contour (i.e. L+H*), and the other to Uncertainty (i.e. L*+H), to avoid confusion or cross-contamination of productions. At each session, participants were initially exposed to a recording of the contour that was the target of that session, produced on a single, monosyllabic target word. A monosyllable was used as a sample to minimize the amount of generalizable information about segmental alignment patterns that could in principle bias the speakers’ own subsequent productions. Participants then read, in exchanges with an experimenter, several practice dialogues of the sort described above, and received ‘corrections’ to productions of non-target contours in the form of additional presentations of the monosyllabic-target sample recording. (Importantly, participants were thereby informed only that their production had strayed from the target pattern; they were not told explicitly how.) All the participants included in the study took to the contours in question quite readily, and rarely deviated from them during elicitation. A surprisingly large number of participants, however, were excluded from the study owing to an inability to reliably reproduce the desired contours in these contexts. For the intended L+H* L-H% Incredulity contour, participants most frequently substituted L* H-H%, the low rising question contour. For the desired L*+H L-H% Uncertainty contour, interestingly, participants most frequently substituted L+H* L-H%. A number of participants simply made this substitution exceptionlessly, and were excluded altogether. Other participants generally produced the desired contours, but deviated periodically from these throughout the elicitations. In these instances, where all four authors were agreed that the contour was not an example of the contours under investigation here, individual productions were discarded from that subject’s data. Productions containing breaks, disfluencies, or insurmountable pitch tracking difficulties were excluded from consideration as well. In order to keep the pool of utterances representing each contour symmetrical with respect to inclusion of target items, whenever a subject’s production of a given target phrase for one contour was excluded for
one of the reasons detailed above, one production of the same phrase from that speaker using the other contour was removed as well.14

Participants read each of the 29 dialogues four times, once in each of two randomized blocks for each contour, for a total of 116 productions per subject. After the exclusion of flawed productions as described above, the total number of productions analyzed was 558. (For the number of productions included from each subject, see Table 1 below). Elicitations took place in a quiet environment, and were recorded directly to laptop at a sampling rate of 44.1 kHz using a Shure Beta 53 omnidirectional headworn condenser microphone and a Sound Devices USBPre digital microphone interface. Subsequent acoustic analysis was conducted using Praat (Boersma and Weenink 2009).

2.1.5. Labelling and analysis

Salient points in the F0 contour and segmental string were labelled in Praat Textgrids. Segmental labels were placed at the beginning of the onset consonant of the pitch-accented syllable, at the beginning of the vowel of the pitch-accented syllable, at the end of the vowel of the pitch-accented syllable, and at the end of the following consonant. Segment boundaries were established through inspection of linked waveform and spectrogram displays in Praat using standard segmentation criteria. (Segmentation was facilitated by the fact that target items had been chosen so that segments in the relevant regions were sequences of alternating vowels and nasals or laterals.)

Two F0 TPs were labelled for each pitch accent, corresponding to the beginning and end of the accentual F0 rise. These TPs, standardly held to reflect the Low and High targets comprising these pitch accents in the literature, were identified as follows: The end of the rise (i.e. the accent’s F0 peak) was labelled simply by locating the F0 maximum within the relevant region. The beginning of the rise was identified as a local minimum in the region preceding the peak. Both points were placed automatically using maximum and minimum detection functions in Praat. This placement was then checked by hand, and where the turning point identified was clearly the product of a pitch tracking error, or a segmental perturbation of the F0 contour, manual correction to the nearest plausible substitute was carried out. (On alternative definitions of these TPs, and in particular the use of F0 ‘elbows’ to mark the beginning of the rise, see note 24 in Section 2.1.7 below.)

Lastly, the location (in time) of TCoG for the Rise-Fall pitch accent in each elicited utterance was calculated in three different ways: using an initial or basic model, and then using two additional models, reflecting certain provisional refinements to the way F0 samples contributing to the derivation of TCoG are collected and weighted in our calculations (on which see Section 2.1.6 below). Figure 5 presents a
fully labelled example of a representative utterance from our corpus, with F0 TPs (including those used as TCoG window edges), segment boundaries, and TCoG marked.

Figure 5. A labelled example showing turning points (I, p, va), segmental markers (oc, bv, bc, ec), and TCoG.

2.1.6. On the derivation of TCoG

The simplest method we have used for locating TCoG corresponds more or less directly to the conceptual description of TCoG given above: F0 is sampled at regular intervals throughout a region of interest beginning at the local F0 minimum corresponding to the onset of the accentual rise, and ending at the F0 minimum representing the transition from fall after peak to phrase-final rise. F0 samples are taken in Hz, with a baseline corresponding to the minimum F0 within the region of interest (in practice, either the beginning of the accentual rise, or the end of the following fall, whichever is lower) subtracted out. TCoG then is just an average of the time points at which samples are taken, weighted by the F0 at each of those points, so that points with higher F0 count more. F0 weighting is done linearly, using measured values in Hz as multipliers.15 Beyond this, no additional weighting factors are applied to sample time points. There is substantial evidence, however, both in the literature, and in our own informal observation, to suggest that this initial model needs revising in two critical areas. Both of these are related to the selection and weighting of individual F0 samples in the derivation of TCoG. The first concerns the nature of the ‘region of interest’, while the second concerns the way in which F0 is used to weight the samples taken therein. These two issues are explored in depth in the following two sections.

2.1.6.1. Weighting samples by their position within the region of interest

Turning first to the region of interest, our initial model defines this window in terms of the F0 contour, with sampling taking place starting at rise onset, and continuing through fall offset. It might be objected at
this point that, since our integration window for TCoG is in practice bounded by $F_0$ TPs, the legitimacy of our claims to a TP-free framework are thereby undermined: If TCoG requires us to locate both the beginning of the rise and the end of the fall, then isn’t our model just as bound to TPs as any other, albeit indirectly? The answer here turns on the extent to which analysis window edges are in fact required to coincide, or even approximate, specific points in the $F_0$ curve. For any single calculation of TCoG for a given contour, some concrete beginning and ending points must clearly be chosen for the analysis window. At least for the data in question, however, where exactly those points are located turns out to matter surprisingly little, in terms of the effect on the location of TCoG, providing only that they land somewhere within a relatively broad lower region to either side of the $F_0$ peak in question. Because low $F_0$ samples included in the analysis window always fall close to the $F_0$ baseline, they contribute comparatively little to the ultimate location of the TCoG. In other words, all TCoG requires is the rough identification of a relatively ‘uneventful’ region between the $F_0$ event in question and any preceding or following events: The window must begin early enough, and end late enough, that the better part of the high $F_0$ region falls within it. Since no particular precision is required here, there is no sense in which the TCoG model actually relies on the precise location of $F_0$ TPs. Our choice of specific TPs to bound the window in the present study is motivated solely by concerns of methodological explicitness and replicability, rather than by any special importance attaching to the TPs themselves. The surprising stability of TCoG location under conditions of such ‘fuzzy’ window demarcation is illustrated in Figure 6 below. Consequences of this stability for the perceptual robustness of TCoG are taken up in Section 2.2.

![Figure 6](image_url)

*Figure 6.* TCoG placement for an L+H* pitch accent, derived using integration windows of five different sizes and orientations, demonstrating the relative stability of TCoG under variation in window placement. Integration window (a) is between local $F_0$ minima. Start and end points for windows (b)-(e) are arbitrarily chosen time stamps.

Ultimately, we believe there are good reasons to prefer an analysis window for TCoG based on something other than the $F_0$ contour itself. Indeed, we have already made substantial progress on an implementation of TCoG using a segmentally-defined analysis window (specifically, a window centered
on the pitch-accented or pitch-accetable syllable). See, in particular Barnes et al. (forthcoming) for some argumentation and details. In the context of the present study, however, we will continue to use the prosodically-defined window described above.

A more serious deficiency with the region of interest as currently defined stems from its categorical treatment of \( F_0 \) samples: samples either fall within the window, in which case they receive the full measure of their \( F_0 \)-modulated weight, or they fall outside it, in which case they count not at all. This, however, turns out to be much too simple. There is much evidence in the literature suggesting that \( F_0 \) contours are more robust perceptually when realized over some regions of the segmental string than over others. For example, the preferential crosslinguistic licensing of tonal contrasts within higher sonority regions (e.g., the syllable rhyme, and more sonorous rhymes in particular) of the segmental string has been attributed to the increased perceptibility of tone during more sonorous segments (Gordon 2001; Zhang 2002 et seq.; Niebuhr 2007b; Barnes et al. 2011a, 2011b; Flemming, to appear). Implicated in particular is the presence of more harmonics, particularly in the lower frequencies, as well as overall signal level. Indeed, earlier models of tone perception that integrate \( F_0 \) information over time have noted that, even over the course of a single vowel, not all portions of an accent-related pitch movement contribute equally to the distillation of a perceived target level for that accent. House (1990), for example, models the endpoints of accentual tone movements by averaging \( F_0 \) over the final 32 ms. of the accented vowel. Likewise, d’Alessandro and colleagues (1998) model the perception of the endpoint of a pitch rise over an isolated synthetic vowel by averaging \( F_0 \) over the final portion of the vowel, with sample weights increasing toward the end of the specified window. Both of these proposals suggest (though for different reasons: spectral stability in House’s case; recency in d’Alessandro’s) the existence of something like a perceptual ‘sweet spot’ within the accented vowel, likely over and above the effects of sonority and signal level referenced above. Further evidence for this kind of asymmetrical sample-weighting is presented by Barnes and colleagues (forthcoming), wherein it is demonstrated that the extent to which sharp-peaked and plateau-shaped pitch accents differ from one another perceptually varies considerably as a function of the way in which the accents under comparison overlap with the accented vowel: if the portions of the two accents in which the shape difference is manifest coincide with the accented vowel, the accents will sound maximally different in terms of timing; if those same difference-bearing portions of the contours happen to fall all or partially outside the accented vowel, the perceptual difference between the accents under comparison decreases, almost to zero in cases where two accents are identical within the accented vowel, differing only after it. Lastly, our own so-far-still-informal observations suggest that, at least for L+H*, changes to contour shape occurring within the rising portion of the accent may have more effect on the perception of tonal timing and scaling than analogous changes.
occurring within the fall, in a manner that, if borne out experimentally, could likewise be related to this kind of asymmetrical sample weighting.  

Before an adequate model of asymmetrical string- or signal-based sample weighting within the region of interest for TCoG can be devised, a great deal more perceptual experimentation, some of it already underway, will need to be completed. This said, we believe certain purely exploratory first steps in the desired direction may still be worthwhile at this time. For example, rather than modulating the weights on samples within the region of interest to correspond to the as-yet-poorly-understood perceptual asymmetries just noted, we might begin our progress in this direction simply by shortening the sampling window for TCoG in such a way as to exclude altogether regions that may contribute less to the percept of tonal timing. For example, since at this stage we are still defining the TCoG sampling window with respect to the F0 contour itself, rather than with respect to the segmental string, as tentatively proposed above, we might restrict this window by sampling only over the region corresponding to the accentual rise, rather than over both the rise and fall, as was done in the initial model. To the extent that the shape of the fall contributes at all to the percept of tonal timing for the accents in question (and evidence from D’Imperio 2000 suggests it does), this step will represent a certain loss of information in the model. We expect an eventual more nuanced method for discounting the perceptual weight of samples from this region to yield improved results over what we report here.  

Calculating TCoG over a window corresponding just to the accentual rise does, however, seem to violate the spirit of our TP-free approach to tonal timing contrasts: We argued above that situating integration window edges in a rough fashion somewhere within the lower regions to either side of our accentual peaks was fundamentally different from, and perceptually more realistic than, extracting the precise locations of F0 TPs within the contour. Defining just the rise, however, would require not just the general locations of the low regions to either side of the peak, but the peak itself. Rather than doing this, and thereby risking the appearance of covert reliance on F0 TPs in the form of a local maximum, for present purposes we will define the region of interest as follows: First, we will locate TCoG for the entire high region as we did above, in a non-TP-reliant fashion. Next, we will use the resulting time point to mark the right edge of a new region of interest. This region of interest, extending from the preceding low up to this original TCoG, is then used for the calculation of a new center of gravity, which we might call TCoG-Rise, to distinguish it from the broader-windowed TCoG implemented in the initial model. The region of interest for TCoG-Rise thus corresponds in a rough way to the accentual rise itself, but again, does not necessitate the extraction of the precise locations of F0 TPs from the signal. While again emphasizing the provisional nature of this revision to the derivation of TCoG, we take the strength of the results reported below to be an indication that some form of asymmetrical weighting of samples within
the region of interest is desirable, and that a more sophisticated model than this one may yield still further improvements.

2.1.6.2. Weighting samples by their F₀ values

The second area in which we believe revisions to the initial model will eventually be required is in the assignment of relative weights to samples based on their measured F₀ values. In the initial model, time points are weighted by their F₀ in such a way that samples with higher F₀ contribute more to the resulting average than samples with lower F₀. This makes sense, given that in the case at hand, it is the timing of a High tone that we are attempting to characterize. Likewise, there is evidence from a variety of sources for something like a greater influence of high F₀ regions over lower ones on judgments of the height of tonal targets during F₀ movements (e.g., Hombert 1975; d’Alessandro, Rosset, and Rossi 1998).

Still, there is reason to believe that the model as described above, in which samples are weighted linearly by the size of the F₀ excursion in Hz over baseline, fails to give sufficient perceptual priority to those higher F₀ samples. For example, in certain instances, TCoG can be inadvertently drawn away from what appears to be the core of the high F₀ region by the combined effect of a large number of lower F₀ samples taken over a longer region. Such a thing is sometimes seen in cases where a rise starts off shallow for a significant duration, continuing that way for a time, followed by a short, steep ascent to the F₀ maximum. Such a contour can be seen in Figure 7. This is an example of an intended (and perceptually unambiguous) instance of L*+H, which is nonetheless miscategorized by the initial model of TCoG as L+H*, owing to a long, gradual rise over the course of the onset consonant and vowel of the accented syllable that, while making little impression perceptually as a rise, nonetheless has the effect of dragging TCoG leftward, away from what is in fact the perceptual heart of the high F₀ region, and into the range of values characteristically associated with L+H*. Analogous mistakes occur in the initial TCoG model for tokens with falls continuing gradually over long, low F₀ regions.
Figure 7. An example of a perceptually unambiguous L*+H miscategorized by the initial model of TCoG as L+H* because of the long low F$_0$ rise during the accented rhyme, which pulls the value of TCoG earlier.

Based on these observations, and following in the footsteps of d’Alessandro and colleagues (1998) in this respect, we believe it is desirable to alter the way weights are assigned by measured F$_0$, such that higher F$_0$ samples are given more influence over the location of TCoG (at least for High accents), while lower samples are given less. As a technical matter, there are many ways this could be accomplished, and we cannot yet advocate with any conviction for one method over another. Our first exploratory attempts at implementing something like non-linear F$_0$ weighting, however, are encouraging enough that they bear reporting here, if only as an indication of a promising direction for the future development of the TCoG model.

The method we investigated for non-linear weighting of samples by their F$_0$ involves transforming measured F$_0$s using a scaled sigmoid function. The sigmoid function appears promising, in that, as depicted in Figure 8, it has the effect both of increasing the weight accorded to samples falling into the upper portion of the pitch movement in question, and decreasing the weight accorded to samples falling into the lower portion.

Figure 8. Illustration of the effect on a simple rising line, with arbitrary values, transformed by the 1 sigmoid.
The particular sigmoid function that we selected is given by Equation 2. These specific parameters provide a sharp ‘on/off’ transition between $F_0$-min and $F_0$-max. The additional scaling factor $N$ was chosen heuristically. The value we used (16) was settled upon by applying a variety of sigmoid transformations to the synthetic $F_0$ contours participants heard during a previous perception experiment. Using these transformed $F_0$ values, we recalculated the location for TCoG in each of these files, and then used the results to model, in a binary logistic regression analysis, the judgments of the participants in that study as to whether the synthetic files represented instances of L+H* or L*+H. We observed in so doing that the predictive capacity of our model initially improved as we increased the scaling factor in the sigmoid transformation (from .5, to 1, 2, 4, and finally 8), whereupon performance at the next two values we tried (16 and 32), appeared to level off. We selected therefore for application to the production data under investigation here the first scaling factor at which this plateau in performance became apparent (i.e. 16).

*Equation 2.* Sigmoid function used to transform $F_0$ values.

$$f_{0_{\text{trans}}} = \frac{100}{998 \left( 1 + e^{-6.9N \left( \frac{f_0-f_{0_{\text{min}}}}{f_{0_{\text{max}}}-f_{0_{\text{min}}}}-1 \right)} \right)} - .001$$

2.1.7. Results and discussion

With these refinements to the derivation of TCoG implemented, we proceed now to a comparison of the relative success of TCoG-alignment and TP-alignment for discrimination of the two contrasting pitch accents under consideration in this study. In order to do this, it is first necessary to select appropriate segmental anchors in the vicinity of the pitch-accented syllable for both Low and High TPs, and for TCoG as well. Typical candidate anchors found throughout the literature include the onset of the pitch-accented vowel, the midpoint of the pitch-accented vowel, the end of the pitch-accented vowel, and the end of the following (ambisyllabic) consonant, as well as some fixed proportion of the pitch-accented vowel, or accented syllable. To determine which candidate was best for each $F_0$ point, analyses were carried out on our data relative to each of these candidate segmental anchors. For each point, only the results for the best candidate anchor are reported below. For the beginning of the rise (hereafter L), this was the onset of the pitch-accented vowel. For the peak (hereafter H), this was the end of the pitch-accented vowel. For TCoG, this turned out to be the midpoint of the pitch-accented vowel.
Beginning first with TCoG, mean distances of TCoG from accented vowel midpoint are given for the contrasting pitch accents in Table 1 below. Data are shown both for each subject individually, and (in the bottom row) pooled across participants.

Table 1. Mean distance from TCoG to midpoint of accented vowel.

<table>
<thead>
<tr>
<th>Subject</th>
<th>L+H* (SD)</th>
<th>L*+H (SD)</th>
<th>Mean Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16 ms. (21, N =44)</td>
<td>81 ms. (26, N =44)</td>
<td>65 ms.</td>
</tr>
<tr>
<td>2</td>
<td>29 ms. (20, N =35)</td>
<td>85 ms. (26, N =35)</td>
<td>56 ms.</td>
</tr>
<tr>
<td>3</td>
<td>31 ms. (16, N =58)</td>
<td>85 ms. (19, N =58)</td>
<td>54 ms.</td>
</tr>
<tr>
<td>4</td>
<td>16 ms. (20, N =30)</td>
<td>90 ms. (26, N =30)</td>
<td>74 ms.</td>
</tr>
<tr>
<td>5</td>
<td>-7 ms. (19, N =54)</td>
<td>104 ms. (28, N =54)</td>
<td>111 ms.</td>
</tr>
<tr>
<td>6</td>
<td>13 ms. (17, N =58)</td>
<td>104 ms. (30, N =58)</td>
<td>91 ms.</td>
</tr>
<tr>
<td>All</td>
<td>16 ms. (23, N =279)</td>
<td>93 ms. (27, N =279)</td>
<td>77 ms.</td>
</tr>
</tbody>
</table>

The substantial separation between TCoG alignments for the two pitch accents achieved for each speaker, together with low within-category variability, suggests that TCoG alignment could function as a reliable cue to pitch accent category. To further probe the strength of TCoG as a cue, we conducted a binary logistic regression analysis, in which TCoG alignment was used to predict actual category membership for the tokens in our elicited corpus. Binary logistic regression is appropriate here because it allows us to assess the relative influence of a number of (in this case continuous) predictors on the value of a single, dichotomous dependent variable (i.e. pitch accent category). The results of this analysis, detailed in Tables 2 and 3 below, make it clear that TCoG achieves significant success in classifying the productions of our participants by pitch accent category. This can be seen, in particular, in the high model chi-square obtained in the Likelihood Ratio test, the high value of Nagelkerke’s $R^2$ and perhaps most transparently in the high overall percent correct shown in the classification table (nearly 95%).

Table 2. Logistic regression results for the TCoG model.

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>SE</th>
<th>Wald $\chi^2$</th>
<th>df</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCoG Alignment</td>
<td>136.590</td>
<td>14.694</td>
<td>86.413</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Constant</td>
<td>-6.831</td>
<td>.721</td>
<td>89.780</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

Note: Model Chi-Square (1) = 605.827, $p < .001$, Nagelkerke’s $R^2$ = .883.

Table 3. Classification table for the TCoG model.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L+H*</td>
<td>L*+H</td>
</tr>
<tr>
<td>L+H*</td>
<td>268</td>
<td>11</td>
</tr>
<tr>
<td>L*+H</td>
<td>21</td>
<td>258</td>
</tr>
<tr>
<td>Overall %</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
At this point we can turn to a comparison of the results of the TCoG-based model with the results of an analogous TP-based model of the same data. Again, this model is based on the alignments of two TPs, corresponding to the L and H target tones in each pitch accent. These were the distance from rise onset to the beginning of the accented vowel, and the distance from peak F0 to the end of the accented vowel. The distributions of these two TPs with respect to their segmental anchors are detailed in Tables 4 and 5 below. Note that while the separation between the means in each case is comparable in size to that between the mean alignments for TCoG reported above, the standard deviations are higher for the TPs. This is particularly so for the rise onset, likely a reflection of the well-known ambiguity of the signal when it comes to identifying Low tones with single, precisely localizable F0 TPs.

Table 4. Mean distance from rise onset (‘L’) to onset of accented vowel.

<table>
<thead>
<tr>
<th>Subject</th>
<th>L+H* (SD)</th>
<th>L*+H (SD)</th>
<th>Mean Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-55 ms. (50, N = 44)</td>
<td>17 ms. (49, N = 44)</td>
<td>72 ms.</td>
</tr>
<tr>
<td>2</td>
<td>-27 ms. (21, N = 35)</td>
<td>23 ms. (36, N = 35)</td>
<td>50 ms.</td>
</tr>
<tr>
<td>3</td>
<td>-51 ms. (29, N = 58)</td>
<td>1 ms. (38, N = 58)</td>
<td>52 ms.</td>
</tr>
<tr>
<td>4</td>
<td>-103 ms. (60, N = 30)</td>
<td>-5 ms. (56, N = 30)</td>
<td>98 ms.</td>
</tr>
<tr>
<td>5</td>
<td>-126 ms. (32, N = 54)</td>
<td>-18 ms. (61, N = 54)</td>
<td>108 ms.</td>
</tr>
<tr>
<td>6</td>
<td>-65 ms. (21, N = 58)</td>
<td>33 ms. (43, N = 58)</td>
<td>98 ms.</td>
</tr>
<tr>
<td>All</td>
<td>-72 ms. (49, N = 279)</td>
<td>8 ms. (51, N = 279)</td>
<td>80 ms.</td>
</tr>
</tbody>
</table>

Table 5. Mean distance from peak F0 (‘H’) to offset of accented vowel.

<table>
<thead>
<tr>
<th>Subject</th>
<th>L+H* (SD)</th>
<th>L*+H (SD)</th>
<th>Mean Separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13 ms. (24, N = 44)</td>
<td>77 ms. (39, N = 44)</td>
<td>64 ms.</td>
</tr>
<tr>
<td>2</td>
<td>22 ms. (39, N = 35)</td>
<td>81 ms. (36, N = 35)</td>
<td>59 ms.</td>
</tr>
<tr>
<td>3</td>
<td>22 ms. (27, N = 58)</td>
<td>76 ms. (26, N = 58)</td>
<td>54 ms.</td>
</tr>
<tr>
<td>4</td>
<td>18 ms. (38, N = 30)</td>
<td>87 ms. (38, N = 30)</td>
<td>69 ms.</td>
</tr>
<tr>
<td>5</td>
<td>3 ms. (26, N = 54)</td>
<td>101 ms. (40, N = 54)</td>
<td>98 ms.</td>
</tr>
<tr>
<td>6</td>
<td>24 ms. (26, N = 58)</td>
<td>97 ms. (34, N = 58)</td>
<td>73 ms.</td>
</tr>
<tr>
<td>All</td>
<td>17 ms. (30, N = 279)</td>
<td>87 ms. (37, N = 279)</td>
<td>70 ms.</td>
</tr>
</tbody>
</table>

Turning to the logistic regression analysis, taken by itself, neither TP quite reaches the level of categorization performance even of the initial TCoG model. (For rise onset by itself: model chi-square (1) = 274.048, p < .001, Nagelkerke’s $R^2 = .517$, correct classification percentage = 77.4%; for peak by itself, model chi-square (1) = 411.744, p < .001, Nagelkerke’s $R^2 = .696$, correct classification percentage =
86.2%, compared with the initial TCoG model’s model chi-square (1) = 482.057, \( p < .001 \), Nagelkerke’s \( R^2 = .771 \), correct classification percentage = 87.8%). When the two TPs are used together as independent variables in a single model, however, matters improve somewhat, as can be seen in Tables 6 and 7.

**Table 6. Logistic regression results for the TP-based model.**

<table>
<thead>
<tr>
<th></th>
<th>( B )</th>
<th>SE</th>
<th>Wald ( \chi^2 )</th>
<th>df</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rise onset alignment</td>
<td>33.881</td>
<td>4.269</td>
<td>62.994</td>
<td>1</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Peak alignment</td>
<td>68.078</td>
<td>7.195</td>
<td>89.517</td>
<td>1</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.204</td>
<td>.352</td>
<td>39.142</td>
<td>1</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Note: Model Chi-Square (1) = 521.262, \( p < .001 \), Nagelkerke’s \( R^2 = .809 \).

**Table 7. Classification table for the TP-based model.**

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>L+H*</td>
<td>258</td>
<td>92.5</td>
</tr>
<tr>
<td>L*+H</td>
<td>30</td>
<td>89.2</td>
</tr>
<tr>
<td>Overall %</td>
<td></td>
<td>90.9</td>
</tr>
</tbody>
</table>

Here we can see significant effects for the alignment of both TPs, and performance for the overall model that is clearly fairly good, though not as good as the refined TCoG-based model just presented. Compare the two, for example, with respect to model chi-square (522 vs. 606, \( p < .001 \) in both cases), Nagelkerke’s \( R^2 \) (.809 vs. .883), and correct classification percentage (90.9% vs. 94.3%).

Another way of conducting this comparison, which makes the same point, would be to include all three potential predictors, TCoG-alignment, rise-onset-alignment, and peak-alignment, within a single, step-wise logistic regression analysis of our data, and to allow the model itself to decide which variables provide significant improvements in fit. Results for the final model produced by such an analysis, using a forward stepwise (likelihood ratio) procedure, are identical to those obtained in the final TCoG model described above: TCoG is selected by the model as a predictor, with the same results reported above, while both rise onset and peak are rejected by the model as not producing a significant improvement in the fit.

To summarize the results of Experiment 1, while the implementational details of a definitive version of the TCoG model will doubtless change as we learn more, we have demonstrated that a TCoG-based model of tonal timing contrasts is capable of outperforming a standard TP-based model in categorization of the pitch accents represented in an extensive sample of carefully controlled, laboratory speech.

Of course, conditions in natural speech are rarely as favorable for the analysis of \( F_0 \) patterns as those found in our database. As reviewed above in Section 1.2, utterances are rarely comprised entirely of sonorants, and even when they are, the maxima and minima, upon which both the TP-based and TCoG approaches implemented above depend, are often quite ambiguous. If we are serious about creating
perceptually realistic models of the phonetics-phonology interface in intonation systems, this is an issue with which our models must at some point come to grips.

As a first step in this direction, the following section presents the results of an additional experiment designed to probe the robustness of the two models, TP-based and TCoG, under circumstances in which the reliability of annotations to the visible F₀ track is considerably diminished. By challenging both models in this way, we are able to gauge how well each might be expected to perform under the less ideal conditions for the extraction of the precise locations of points within the F₀ curve that are typical of natural speech. Additionally, we are able to test a central claim of the TCoG model—namely, that the derivation of TCoG does not in fact require the location of precise points within the F₀ curve, and thus that the TCoG model remains distinct from the TP-based models that preceded it.

2.2. Experiment 2: Robustness of categorization under adverse conditions for F₀ analysis

We have argued above that the extraction of the precise locations of specific points in the F₀ curve in perception is for a number of reasons a dubious proposition: In many instances, would-be TPs are either irretrievable (owing to, e.g., spans of voicelessness or irregular pulsing) or difficult to locate unambiguously (in high plateaux, extended low troughs, or shallow rises and falls). Furthermore, even under ideal circumstances, evidence from the psychoacoustic literature suggests that human listeners are simply not particularly good at tasks such as, e.g., estimating the heights of the onsets and offsets of pitch movements realized over speech-like stimuli (Rossi 1971, 1978; d’Alessandro, Rosset, and Rossi 1998).

We have further argued that, unlike TP-based models, the TCoG model is less susceptible to problems of this sort because, while in practical terms the window over which TCoG is calculated must begin and end somewhere, it appears unnecessary that it should begin and end anywhere in particular, provided its edges are set somewhere within a relatively broad ‘uneventful’ region to either side of the F₀ event in question. That is, we hypothesized that TCoG is robust to ‘noisy’ placement of window edges in a way that TP-based models would not be, vis-à-vis the TPs in question. The following experiment is designed to test this hypothesis.

2.2.1. Methods

This experiment makes use of the same corpus of 558 productions of rise-fall-rise contours elicited from six speakers as part of Experiment 1. (See above for details of language materials and elicitation techniques.) Experiment 2 likewise involves the same kind of analysis in terms of classification via logistic regression that were presented for these data in Experiment 1. Now, however, we sought to add a
degree of random variability to the classification process by re-running those analyses, this time using F₀ points (that is, TPs and TCoG window edges) the labels for which had been shifted in time along the F₀ curve by randomly determined amounts. Since extreme perturbation of the locations of these labels would likely disrupt classification of contours in any model, we decided to begin with a level of added ‘noise’ designed to produce relatively minor perturbations to the positions of our original F₀ point labels, and then to gradually increase the level of added labelling noise in subsequent analyses.

Relabelling for this experiment was accomplished as follows: for every utterance in the corpus, each F₀ label was displaced in time by a random number of milliseconds, selected from a Gaussian distribution with a mean of zero, and a standard deviation equal to some multiple of each speaker’s actual standard deviation for the realization of the relevant F₀ point within each contour. At the first level of additive noise, this multiple was .5. Thus, for example, if the standard deviation of a given speaker’s peak alignments with respect to the end of the pitch-accented vowel for an L+H* pitch accent was 18 ms., then each label indicating an L+H* peak in that speaker’s productions was moved by a number of milliseconds randomly selected from a Gaussian distribution with a mean of zero and a standard deviation of 9. Selecting, e.g., a value of 5 from that distribution as a displacement magnitude for a given peak label would then require moving the relevant label 5 milliseconds later along the F₀ track from its original position. A value of -5, by contrast, would involve shifting the label earlier by the same amount. (To be clear, at issue in all cases are only the labels identifying the relevant points within each F₀ curve. The estimated F₀ curves themselves were not altered in any way.) The labels that were perturbed in this fashion were: the F₀ minimum preceding the accentual rise (used to represent both the rise onset TP, and the left edge of the TCoG integration window), the F₀ maximum within the high region, and the F₀ minimum following the fall from the accent’s peak (used as a right edge for the TCoG window).

Such relabelled versions of the data set were created using seven different levels of added noise, representing random distributions with standard deviations of .5, 1, 2, 3, 5, 8, and 10 times speakers’ actual standard deviations for the alignment of each F₀ point respectively. To avoid the creation of labelled contours violating principles of common sense, however, certain constraints were placed on the output of the relabelling procedure. These were the following: 1) perturbed rise onset labels and TCoG window beginnings were not permitted to follow the perturbed label for peak, and 2) perturbed TCoG window ends were not permitted to precede perturbed peak labels. Additionally, for practical reasons relating to F₀ track estimation, no labelled point was permitted to occur outside the target phrase itself (e.g., during the preceding word ‘there’s’, or in the region following the last reliable F₀ estimate in the final word of the utterance). In practice, since none of the original, noise-free labels occurred outside this region either, this seemed like a reasonable condition to impose. In each case, if the output of the noise
generation procedure for a given utterance violated one of these constraints, the random number selection for all points was rerun until an output that satisfied all constraints was achieved.

Upon completion of relabelling at each level of Gaussian noise, new TCoGs were calculated for each utterance using the new, noisy F₀ window edge labels, and these, together with new, noisy TP alignments, were submitted to logistic regression analyses for classification as described in Experiment 1. In order to assure that an adequate picture of each model’s performance at each noise level was achieved, the noise-generation and subsequent relabelling and reanalysis procedure was carried out a total of 30 times at each noise level.

2.2.2. Results

Owing to the large number of distinct logistic regression analyses conducted here (2 models X 7 noise levels X 30 trials = 420), complete details of each regression model will not be given here, but instead, results will be characterized only in terms of percent correct classifications (as described above for Experiment 1). Figure 9 shows the results of these noisy classification analyses for the two-TP (rise onset and peak) and TCoG models, represented in terms of mean percent correct for each model over the 30 trials conducted at each level of random noise. (Error bars represent standard deviations.) The pattern here is quite clear: while the performance of both models eventually degrades to some extent with the addition of labelling noise, the difference between the two models present in the original analysis is quickly exaggerated, as the accuracy of the TP-based model degrades far more rapidly than does that of the TCoG model with each added level of Gaussian noise. Furthermore, while the TCoG model appears to level off in performance around 85% correct in the noisiest conditions, the Two-TP model drops all the way to the mid-60s, and may still be declining gradually by the time the noisiest level is reached.

This outcome confirms the hypothesis that a TP-based model of pitch accent discrimination requires a comparatively high level of precision in the locations of its F₀ TPs in order to be successful. In contrast, a TCoG-based analysis requires far less precision than this in the placement of its analysis window edges: as long as the region of interest begins and ends within a relatively broad region to either side of the F₀ peak, the location of TCoG remains remarkably stable. Derivation of TCoG can thus be said not to rely on the extraction of specific, precisely located TPs, and for this reason is both more robust and perceptually more realistic than a comparable TP-based alternative. All that is necessary for it to function is an estimate, however rough and approximate, of the beginning and end of the high F₀ region realizing the pitch accent in question.
Figure 9. Mean percent correct classification of L+H* and L*+H tokens with 8 levels of ‘noise’ added to the estimates of TP and TCoG window edge locations, for the TCoG model (top curve) vs. the Two-TP-based model (bottom curve). The performance of the Two-TP-based model falls off much more rapidly as labelling noise increases, indicating the robustness of the TCoG model to ambiguities in turning point locations.

3. Contour shape in the perception of tonal timing: An overview and a reappraisal

In the preceding, we have shown that TCoG-based models are capable of outperforming standard TP-based models of tonal implementation in the task of classifying pitch accents by tonal timing pattern for utterances produced under the controlled conditions of laboratory speech. We have also demonstrated that TCoG is more robust to the kinds of ambiguities that arise in the analysis of F0 contours in natural speech, because unlike TP-based models, TCoG does not require the precise location of any particular point or points within the F0 contour, and thus remains relatively stable under conditions of variability in the positioning of its analysis window.

Beyond these practical virtues, however, we also argued in the introduction that TCoG is capable of capturing in a unified, explanatory fashion a range of perceptual effects involving global contour shape that remain puzzling, both collectively and individually, from the point of view of F0 TPs. One of the central virtues of the TCoG model is that it allows us to see a common principle at work here, in what otherwise appear to be unrelated quirks of tone perception. In Section 1.3 above, we demonstrated how one of these effects, involving the asymmetrical scaling of rise onsets and fall onsets, is accounted for by the TCoG framework in terms of left/rightward shift of the bulk of the high F0 region. In the following sections, we show how this same account can be extended to all remaining effects in the literature of which we are currently aware.
3.1. Effect of pitch movement curvature

Aspects of the curvature of F₀ rises and falls, often characterized informally in terms of ‘scooped’ or ‘domed’ pitch movements, has been invoked in the description of contrastive elements in the intonation systems of a variety of languages. In English, for example, the rise taking place during the implementation of the L*+H pitch accent is often described as ‘scooped’. Likewise, Dombrowski & Niebuhr (2005) for German, and Welby (2003) for French both invoke characteristic pitch movement shapes as associated with particular phonological categories. Barnes and colleagues (2010b) have taken up the topic of pitch movement curvature from the point of view of its effect on the perception of tonal timing contrasts. If it is true, for example, that L*+H, with its later pattern of tonal timing, is characteristically realized with a more concave or ‘scoopy’ rise shape than is the earlier-timed L+H*, we might expect that there is a reason that the association of shape and timing should run in this direction and not, for example, the reverse. It might be, in other words, that pitch movement curvature is being used in this instance to enhance the perceptual salience of the tonal timing contrast between L*+H and L+H*.

This is in fact precisely what we would expect from the point of view of TCoG. As Figure 10 below makes clear, in comparison with a perfectly symmetrical linear rise-fall pattern, a pitch accent realized with a scoopier rise will have a relatively later TCoG, owing to the rightward shift of the bulk of the High F₀ region. Likewise, a domier rise in the same context would yield an earlier TCoG. We predict, then, that in perception scoooper rises should bias listeners toward judgments of later tonal timing, just as did right-weighted rise-falls in D’Imperio (2000), while domier rises should bias listeners toward judgments of earlier tonal timing, just as did D’Imperio’s left-weighted rise-falls. Barnes and colleagues (2010b) demonstrate the correctness of both these predictions in a perceptual experiment inspired by D’Imperio’s studies.

![Figure 10. Schematic illustrating the effect on perceived alignment of TCoG of a ‘scooper’ (left panel) or ‘domier’ (right panel) rise shape for a Rise-Fall pitch accent.](image)

To reiterate, at first glance, it might not be obvious what rise/fall scaling asymmetries of the kind investigated by D’Imperio should have to do with the curvature of individual pitch movements, or why
either of these matters should have ramifications for the perception of tonal timing contrasts. TCoG makes it clear what the connection is, and predicts precisely the effects documented in the literature.

3.2. Effect of pitch movement durations

The effects of F₀ movement durations (and thus also slopes) on listener categorizations of pitch accent alignment have been investigated in detail in a number of recent studies. For example, Niebuhr (2007a) investigates the perception of a three-way accent alignment contrast in German, the subject of substantial perceptual research since Kohler (1987). Niebuhr’s study, similar to those of D’Imperio (2000), replicates Kohler’s earlier finding that by shifting an accentual rise-fall F₀ pattern (that is, by moving all relevant turning points at once) through an artificial continuum of alignments with respect to the accented syllable, relatively sharp perceptual boundaries can be seen to emerge between alignments judged to represent the German ‘early peak’, ‘medial peak’, and ‘late peak’ pitch accent categories. Niebuhr goes on to demonstrate, however, that for a given alignment of F₀ peak, function-based judgments of pitch accent identity could be shifted across categories by changes to peak ‘shape’ involving the lengthening and shortening of rises and falls in various combinations. Generalizing broadly, what Niebuhr shows is a tendency, under the right circumstances, for accents with longer, more gradual falls to skew listener responses toward late alignment, while accents with gradual rises and sharper falls tended to cue the percept of earlier alignment, despite identical peak positions. These results mirror similar findings of D’Imperio (2000), who also manipulated rise and fall durations. Like Niebuhr, D’Imperio finds that, at least at certain points in her continua, manipulating the durations of Neapolitan Italian accentual rises and falls can shift listener judgments to earlier or later alignment types. The direction of the effect, furthermore, is the same as that found by Niebuhr. The basic pattern here is illustrated in Figure 11.

![Figure 11. Schematic illustration of the effect on TCoG alignment of extending the duration of the fall (left panel) or rise (right panel) of an accent-related rise-fall F₀ contour.](image)

This finding is not in itself troubling for a TP-based approach to intonational contrasts; extending the durations of falls and rises while leaving peaks in place does involve changes to the alignment of F₀ TPs,
and thus is not in principle beyond the descriptive capabilities of the TP-based approach. What is a problem, however, is the set of TPs that seem to be endowed with this potential influence over listener categorizations. D’Imperio, for example, demonstrates that within a continuum of alignments for a rise-fall accent shape ranging between what are analyzed phonologically in Neapolitan as an earlier-timed L+H* pitch accent, and a later-timed L*+H, participants’ propensity to render ‘late alignment’ judgments (that is, L*+H) can be increased by lengthening the duration of the fall following the accentual peak. In this case, however, the TP corresponding to the end of the fall does not represent the realization of either of the two tonal targets constituting the pitch accents in question. The role this TP seems to play, therefore, in determining listeners’ categorizations of these pitch accents, becomes somewhat mysterious.

Similarly, Niebuhr demonstrates in his study of timing contrasts in German that for a contrast between an early-aligned and a medially-aligned High pitch accent, the percentage of ‘early’ responses could be increased by lengthening the duration of the rise to the accent’s peak, despite the fact that, in GToBI terms at least, the two competing accents would be analyzed as H+L* and H* respectively.28 Once again, then, the problem is that the TP whose manipulation appears to influence categorization of the pitch accent is not one standardly held to reflect any tonal target included in the representation of the accent in question.29

In a TCoG-based approach, these results again fall out naturally: assuming a constant peak height, when the duration of the fall from a pitch accent increases, it naturally increases the area under the curve of the portion of the high F0 region following the peak. This shifts the TCoG rightward, which we predict would bias listeners toward judgments of later accent alignments, regardless of whether those judgments involve shifts from H*+L to H*, or from H* to L*+H in ToBI terms. Likewise, increased rise durations, again regardless of the pitch accent in question, should shift the TCoG of the high region leftward, and so should make the contour’s alignment sound earlier, just as demonstrated by Niebuhr and D’Imperio.30

3.3. Effect of peak shape

A final phenomenon relevant to this investigation concerns differences in the perception of tonal timing patterns associated with F0 peaks of differing shapes. Differences in the perceived scaling of pitch accents implemented as sharp peaks, and those realized with an extended region around the maximum (known as ‘plateaux’) are well-documented (’t Hart 1991; D’Imperio 2000; Knight 2003, 2008): all things being equal, a plateau-shaped accent sounds higher in pitch to listeners than an accent with sharp peaks at the identical maximum F0. This result already suggests something more complex occurring in perception than the simple extraction of the location in F0 space of a single TP. Additionally, however, the difference between sharp peaks and plateaux has also been shown to have an effect on the perception of tonal timing.
contrasts. D’Imperio (2000), for example, demonstrates that, all things being equal, contours rising to a plateau beginning at a given point in a synthetic alignment continuum are associated with significantly more ‘late’ alignment judgments on the part of listeners than contours with an instantaneous peak at that same time point. If the beginning of a plateau is not equivalent to a sharp peak for judgments of alignment, however, neither is the end of the plateau. A peak occurring at a given step in D’Imperio’s continua elicited more late-alignment judgments than a plateau ending at the same location. This is schematized in Figure 12.

![Figure 12](image_url)

*Figure 12. Schematic illustration of the effect of peak shape on perceived alignment of a Rise-Fall pitch accent in D’Imperio’s (2000) study of Neapolitan Italian. The left panel shows that perception shifts later when the plateau (solid line) extends later than the corresponding peak (dotted line), while the right panel shows an analogous perceptual shift earlier when the plateau begins earlier than the corresponding peak.*

Again, it is not clear how these results are to be understood in terms of F_0 TPs. Somehow, the location of the target associated with the High tone seems to be at issue, and yet within the high, flat region corresponding to the plateau, there is no single TP that can be identified with the changes in perception documented by D’Imperio. In terms of TCoG, however, this result seems less mysterious: Since, in the case of an F_0 plateau, the bulk of the high F_0 region (and thus the TCoG) falls in between the end of the rise and the beginning of the fall, we would predict that all things being equal, an F_0 plateau should sound later than a sharp peak aligned at plateau onset, but earlier than a sharp peak aligned at plateau offset. This is essentially what D’Imperio reports.  

3.4. Summary of shape effects on perceived F_0 alignment

Figure 13 presents a synthesis of the contour shape phenomena described in the preceding sections. Two things should be clear at this point: From the point of view of F_0 TPs, these patterns are difficult to account for even individually; taken collectively, the resulting pattern is all the more unexpected. TCoG, on the other hand, affords us a way of understanding what this seemingly unrelated group of contour shape phenomena has in common: For each curve, what matters is where the TCoG is located relative to salient points in the segmental string. Changes to contour shape that result in shifts of the bulk of the high
region rightward relative to these points result in a perceptual bias toward phonological categories with later tonal timing patterns, while changes that shift the bulk earlier, however disparately they may achieve this, will in the same way bias listeners toward categories with earlier tonal timing patterns.

Figure 13. Summary chart: left-weighting and right-weighting shape modifications. Solid lines show changes in $F_0$ shape and vertical black lines show resulting change in perceived alignment.

Beyond this, however, TCoG makes typological predictions regarding the kinds of alignment and contour shape patterns we should expect to find co-occurring in the languages of the world. For example, to the extent that scoopier rises enhance the percept of later high-tone timing, we might expect to find the combination of scoopy rises and later alignment co-occurring in tone systems with unusual frequency. Likewise, we might predict that languages would avoid realizing contrastively later-timed High tones with domier rise shapes, because this combination is predicted to be perceptually maladaptive. In the same fashion, we might expect to find languages preferring, all things being equal, to use the relative scaling of rise onsets and fall offsets to right-weight High tones with later characteristic timing patterns,
but to left-weight Highs with earlier timing. The opposite pattern (left-weighted late accents and right-weighted earlies), we would predict, should be less robustly attested. It is too early to know whether these patterns are indeed observed cross-linguistically. It is worth pointing out, however, that precisely this pattern of left/right-weighting of pitch accents appears to obtain for the contrast between L+H* and L*+H in Neapolitan Italian described by D’Imperio. This contrast is characterized, D’Imperio shows, by later alignments for rise onset, peak, and rise offset in the case of L*+H than in the case of L+H*. There is also a strong tendency for F0 at the beginning of the rise in Neapolitan to be lower in L*+H than in L+H*, and for the end of the fall to be higher in L*+H than in L+H*. This configuration, depicted in Figures 14 and 15 (reproduced from D’Imperio 2000), is just what the TCoG hypothesis predicts: L*+H has its later timing pattern enhanced perceptually by right-weighting, while L+H* has its contrastively earlier timing pattern enhanced by left-weighting.

Figure 14. F0 traces for question and statement versions of the Neapolitan Italian sentence Mama andava a ballare a Lalla, "Mom used to go dancing at Lalla’s" (From D’Imperio 2000: 2). Vertical line marks the onset of the stressed initial vowel of Lalla.
3.5. TCoG in phonetics and phonology: A comparison with existing models

In Section 1.3 above, we stressed the fact that the TCoG model is not a repudiation of the findings or insights of the TP-based segmental alignment research program. There can now be little doubt that speakers are systematic in the way they time the onsets and offsets of their F₀ movements, and clearly the timing of those onsets and offsets can influence perception in a variety of ways. Observed systematicity alone, however, does not allow us to conclude that the goal of TP-alignment by speakers in production is simply the recovery of the precise locations of said TPs by listeners in perception. Rather, we suspect that while systematic alignment of TPs may be a demonstrable fact about speech production, it is ultimately a strategy that serves a larger end in speech perception, namely, determining the alignment of TCoG with respect to the accented syllable.

We also stressed the compatibility of TCoG with the fundamental principles of the AM approach to intonational phonology, suggesting in Section 1.3 that TCoG might best be understood as an auditory target for the implementation of tone-level specifications within a model like AM. Under this approach, each tone specification would receive its own target TCoG localization in time and frequency space. Indeed, to the extent that TCoG involves the perceptual integration of multiple aspects of the acoustic signal, it may prove fruitful to consider TCoG as something like the Intermediate Perceptual Property (IPP) of Kingston and Diehl (1995, et seq.). IPPs for these authors represent the perceptual integration of a range of auditorily similar features of the acoustic signal, and thus occupy a position in between the signal itself, and the more abstract level of the distinctive feature, which they, like many, take to be the
substance of phonological representations. A given distinctive feature (e.g., [voice]) may be cued by one or more IPPs (e.g., the ‘low-frequency property’ and the ‘C/V duration ratio’), each of which involves the integration of a range of acoustic properties of the signal (e.g., closure voicing, low F1, and low F0 in vowels flanking voiced stops). In the case of TCoG, then, we could understand a particular constellation of facts about gestural timing, shape, and scaling as integrating to produce an IPP (such as ‘early TcoG’, or ‘late TcoG’), which itself would function, potentially along with other IPPs, to cue a particular phonological category (e.g., L+H*).

It should also be stressed, likewise, that TCoG is not by itself intended to serve as a fully specified model of tone production, in that a given TCoG target location in time and F0 space critically underdetermines what a speaker will actually do (i.e. which combination of alignment and shape characteristics he or she will use) to achieve that target. What TCoG supplies instead is an auditory target, along with explicit constraints on the kinds of shape and timing combinations that could succeed in reaching that target. Which combinations are in fact selected by a given community or individual may be somewhat arbitrary, and substantial within-category variation is predicted to be possible without detriment to tonal identification. (See Niebuhr, et al. 2011, for the possibility that speakers of a given language in fact fall into two camps, ‘shapers’ and ‘aligners’, in terms of how they realize tonal targets.)

Within the space of theoretically possible combinations of shape and alignment that would achieve a given target TCoG location, we furthermore expect speakers’ actual production options to be limited by a range of additional constraints, unrelated to TCoG. For example, Cho (2010) provides evidence from a number of languages for a model of tonal implementation based on functionally grounded, weighted constraints. One set of constraints concerns pitch movement shape, primarily construed as rise slope. Rising accents, for example, might be optimally realized with slopes within a particular window (where too steep a slope would be overly effortful, and too shallow might be insufficiently robust). Additionally, a substantial and growing body of literature coming from the context of Articulatory Phonology (Browman and Goldstein 1986, 1992) is focused on the notion that what TPs are in fact aligning themselves with in production is not acoustic but articulatorily landmarks; in other words other turning points, such as the onsets and offsets of segmental gestures in speech production (Ladd 2006; Xu and Liu 2006; D’Imperio et al. 2007; Mücke et al. 2009; Niemann et al. 2011, inter alia). Attested TP-alignment patterns might then be seen to result both from perceptually-oriented TCoG-based constraints on timing and contour shape, plus constraints on intergestural coordination, mandating that the choice of alignment patterns be from among a small set of articulatorily favorable synchronizations. Crucially, however, TP-alignment would remain just one (if a particularly efficient one) among a variety of strategies available to the speaker for the achievement of the auditory target in question.
We take as evidence for this conclusion the fact that, in English at least, robust categorization of $L^+H^*$ and $L^*+H$ pitch accents seems not to be obtainable using the alignment of either the beginning of the rise (i.e. ‘L’) or the $F_0$ maximum (i.e. ‘H’) alone, but rather only by including the locations of both turning points in the regression model presented in Experiment 1 above. The fact that for a ‘delayed peak’ accent, both the beginning of the rise and the peak are realized systematically later suggests that the two turning points are working together toward a common perceptual goal. If what actually mattered were just the locations of each turning point, rather than a perceptual gestalt arising (in part) from the locations of both turning points, as we are arguing, then there would be no particular reason for them to move in unison: in principle at least (and subject of course to constraints on maximum speeds of pitch movements and the like), we might equally expect to find contrasts between, for example, a pitch accent with an early rise and a late peak on the one hand, and a pitch accent with a late rise and an early peak on the other (or indeed, any other arbitrarily selected pairing of rise and peak alignment patterns, so long as those alignments were systematic). To the extent that we do not appear to find such freedom in the constitution of alignment contrasts, we suggest that this is because the goal in realizing accentual alignment contrasts is the synchronous manipulation of multiple turning points, together with other aspects of contour shape, to produce a particular TCoG profile for the accent in question. This view predicts that certain combinations of shape and TP-alignment should be dispreferred among the languages of the world as the basis for phonological contrasts.

Lastly, while most of the preceding discussion has contrasted TCoG specifically with approaches to phonetic implementation within AM phonological models, we can also compare TCoG with other approaches to the phonetics of intonation. In particular, it is important to differentiate what TCoG is meant to accomplish from what other existing models of contour shape both aim at and succeed in doing. Approaches such as the Fujisaki model (Fujisaki and Hirose 1984), Momel (Hirst and Espesser 1993), and Tilt (Taylor 2000), for example, all encode contour shape in one form or another, but differ from TCoG in important ways. The Fujisaki model and Momel are concerned with representing, with as little deviation as possible, the precise shape of each individual $F_0$ contour. Their representations may be considered as something like lower-dimensional compressions of the raw contour itself. They make no predictions, as we understand them, about the relative perceptual salience of various shape modifications, about the interaction of different aspects of contour shape with one another or with timing, or about the mapping of these things to phonological categories. Likewise, they do not aim to predict possible and impossible contrast types, or phonological inventories. TCoG makes strong predictions on all these fronts. It also differs from those approaches in that it fails, by design, to represent the contour in such a way that one could map successfully from a single representation of the contour in terms of TCoG back to a unique description of that contour itself. Indeed, this one-to-many mapping between TCoG and various
combinations of pitch movement shape and timing is what allows TCoG to make the predictions regarding the phonetics/phonology interface that it does.

Among the phonetic models of contour shape, the one possible exception to the preceding is the Tilt model, which, as noted above, does share certain goals and insights with TCoG. Like Momel and the Fujisaki model, Tilt focuses on deriving an accurate, lower-dimensional representation of the raw contour itself with as little information loss as possible. Information loss is quantified in Taylor (2000) as simply the RMS error between the contour generated by Tilt and the original contour. Again, this is explicitly not the kind of accuracy to which TCoG aspires. On the other hand, Tilt also aspires to achieve a representation that encodes only those aspects of contour shape that are linguistically significant. In this sense, Tilt and TCoG share a certain spirit of purpose: Corresponding to the three acoustic parameters that Tilt uses to characterize the shape of F0 events (timing being separate entirely), it also posits a set of three analogous continuously-valued scales in phonological representation, similar in function to the traditional discrete distinctive features of structuralist phonology. Of these, Taylor says, “The key point about the scales of phonological representation is that events which are perceived as being the same should have the same values in the Tilt representation.”

This is essentially the position we have been advocating here, with a key difference. We have shown that in order to express the functional unity of the full range of shape and timing combinations that may be ‘perceived as being the same’, or linguistically equivalent, it is necessary to adopt a still lower-dimensional representation than the one formalized in Tilt. Tilt’s three parameters, from which timing stands apart, are the amplitude (or magnitude) of an F0 event, the duration of the event, and the Tilt parameter (a combined expression of what is essentially event lopsidedness in both the frequency and time domains). Like TCoG, the Tilt parameter successfully expresses the functional unity of what we have impressionistically called left/right-weighting in the scaling dimension, and asymmetrical rise/fall duration in the temporal. Taylor argues for the integration of these aspects of shape into a single parameter on the grounds that they covary in speech, while we have argued the same primarily on the basis of their perceptual equivalence. Beyond this, however, the parameters are allowed to combine freely, and no predictions are made regarding the effects on perception of their combination either with one another, or with event timing patterns. Thus in Tilt it is not clear why right-weighting of a rise-fall accent in the scaling dimension appears to have the same effect on perception as shifting the entire fall later in time to create a plateau. (The former would involve the tilt parameter, while the latter would involve event duration.) Or again, why the same effect on the perception of timing could be achieved by simply moving the entire shape, unaltered, later with respect to the segmental skeleton.

TCoG accounts for these perceptual effects, and more, while also making a range of explicit predictions about what kinds of shape/timing combinations should be most effective in signalling a given
event type, and what therefore should be common crosslinguistically. The way it does this is by reducing
the number of parameters characterizing $F_0$ events further still, to just two: the location of TCoG in time
(the topic of this paper) and the scaling of TCoG in frequency space. As with Tilt, the mapping to
phonology is thus also one-to-one, if we assume that phonological representations contain just
information about tonal timing (TBU identity), and tone scaling (i.e. H vs. L). There is a cost, however,
associated with the greater explanatory coverage achieved by TCoG: In lowering the dimensionality of
the description to account for the effects we observe, it becomes necessary to relinquish the possibility of
accurately generating a unique, fully-specified $F_0$ contour from its TCoG representation alone. Instead, we
move to the kind of one-to-many mapping described above. We have argued that this is desirable as an
explanatory device in phonetic theory. As a compression algorithm allowing faithful recovery of the
original $F_0$ contour, of course, it is less than ideal.

4. Conclusions

The aim of the foregoing has been to present evidence in favor of a new model of the
phonetics/phonology interface in intonation systems. This model, which we have called Tonal Center of
Gravity, approaches tonal implementation in both time and $F_0$ space not in terms of the precise location of
any particular identifiable point or points within the $F_0$ contour, but instead by focusing on how the bulk
or mass of a given $F_0$ event is distributed within those two dimensions. In this paper, we have
concentrated exclusively on the time dimension, and limited ourselves to discussion of the realization of
two of the pitch accents of English containing High tonal targets. Within this domain we have
demonstrated that TCoG-based models are capable of outperforming TP-based models in categorizing
contextualized productions of American English pitch accents characterized by contrasting tonal timing
patterns (Experiment 1). We have also demonstrated the greater robustness of the TCoG-based model to
variability in the analysis of the $F_0$ contour itself: because it does not rely on the precise locations of
particular $F_0$ points, TCoG can be reliably located for a given $F_0$ event even where such traditional
landmarks as the onsets or offsets of constituent $F_0$ movements cannot be located unambiguously
(Experiment 2). Many problems with the use of $F_0$ TPs as the sole characterization of $F_0$ events in natural
speech are thus avoided with the adoption of this model. Lastly, TCoG provides a unified, explanatory
account of a broad and seemingly disparate set of phenomena whereby global contour shape is seen to
influence the perception of tonal timing patterns in ways that $F_0$ TPs are ill-suited to account for. Precisely
located $F_0$ TPs, in other words, are neither necessary nor sufficient to account for the mapping between
phonological categories in intonation systems and observable $F_0$ curves.
TCoG succeeds in capturing key insights of configuration-based approaches to intonation systems, while at the same time avoiding the inclusion of contour shape per se within phonological representations of F0 events. It thus retains what we see as the clear advantages of a level-based approach to intonational phonology. Along the way, TCoG also makes a number of unexpected predictions about the effects of F0 contour shape on perception that are confirmed by the existing literature, as well as predictions regarding the typology of tone systems in the languages of the world. The next steps in the TCoG research program must therefore be to investigate these predictions further.

Appendix: A complete list of phrases produced by participants in Experiment 1, organized by type

1. Two syllables, first syllable stressed:
   lemon, melon, money, nanny

2. Three syllables, first syllable stressed:
   lemony, melony, minimum, nominal

3. Four syllables, first syllable stressed:
   lemonier, melonier, minimally, nominally

4. Four syllables, main stress on first syllable, secondary on last:
   laminalize, mineralize, nominalize, minimalize, mimeograph

5. Four syllables, main stress on first syllable, secondary on penult:
   laminated, luminary, millimeter, numerator

6. Two-word compounds, with disyllabic first word receiving main stress on first syllable, and the final syllable of the disyllabic second word receiving secondary stress (stress patterns analogous to those of type 4):
   melon alert, money lament, nanny morale, lemon alarm

7. Two-word compounds, with monosyllabic first word receiving main stress, and penult of the 3-syllable second word receiving secondary stress (stress patterns analogous to those of type 5):
   knee maneuver, law marauder, lime aroma, nail enamel

Notes

1 Arvaniti & Ladd (2009), for example, liken the relationship between tones and TPs to that between an F2 maximum in the spectral profile of the word Maya, and the phonological segment /j/ that gives rise to it.

2 This seems to be the thinking, for example, behind aspects of the Momel (Hirst & Espesser 1993, et seq.) algorithm for the stylisation of F0 contours. Momel uses quadratic spline fitting to arrive at a
‘phonetic representation’ of intonation contours. The output of Momel stylisation is a smooth, continuous curve, projected even through voiceless regions, with ‘target points’ (i.e. TPs) identified throughout.

3 In fact, the judgment they were making was between question and statement interpretations of the same sentence, a distinction that, in this dialect of Italian, is carried by the contrast between the pitch accents analyzed by D’Imperio as L+H* and L*+H.

4 Indeed, these results lead D’Imperio to question the role of F0 turning points in the perception of alignment contrasts such as this one, suggesting instead that what listeners may in fact be doing involves the formation of a ‘gestalt’-like percept of accent location, to which not only TP-alignments, but also other factors grouped rather loosely under the rubric of contour shape might contribute. As will be seen in the following sections, we think this is exactly right.

5 It is therefore a relative of earlier work in this vein approaching tonal structure from the point of view of area under the F0 curve (Segerup and Nolan 2006; Knight 2008; Barnes et al. 2010a). TCoG is different, however, in that it does not involve the literal accumulation of F0 ‘mass’ over time, and therefore avoids certain pathological predictions to which approaches based on raw AUC are prone (e.g., that falls and rises with identical AUC over the same region should sound equivalent, or that 100 ms of [a] at 200 Hz should be equivalent to 200 ms of [a] at 100 Hz).

6 The idea that a gestalt measure such as center of gravity might be useful in characterizing the reference location of an object turns out to be a familiar one from the literature on object localization in visual perception. In particular, studies of saccadic localization of objects have shown that for spatially extended targets, saccades land with a high degree of consistency at the center of gravity of the object in question, even when the center of gravity is located outside the boundaries of the object itself. (See, e.g., Kovács 1996; Melcher and Kowler 1999; Vishwanath and Kowler, 2003.) Without overstating the parallels, we find the possibility of convergent results in the study of visual object localization and auditory event localization exciting indeed.

7 In this initial implementation of TCoG, we include only sample F0 as a weighting factor. Ultimately, it will be necessary to include additional weighting factors as well, corresponding to the variety of additional acoustic parameters known to influence the perception of pitch, e.g., segment intensity, or some derivate thereof (Niebuhr 2007b, Barnes, et al. 2011a, 2011b).

8 The attentive reader may at this point recognize a similarity in what TCoG accomplishes to what the Tilt parameter of Taylor (2000) was meant to achieve. For more on Tilt, see Section 3.5 below.

9 Here, Willems and colleagues’ (1988) distinction between perceptual tolerance and linguistic tolerance in F0 contour stylization becomes useful to invoke. Our claim is not that these contours would necessarily be perceptually indistinguishable, but rather just that they would remain functionally equivalent, in the sense that they would be judged to be well-formed instances of the same linguistic category.

10 As should be clear from the examples given above, the magnitude of shape effects on the location of TCoG is typically relatively small in comparison with the magnitude of the differences that can be achieved by shifting the ‘anchor points’ of the pitch movements in question. In this sense, from an articulatory point of view, TP-alignment may prove to be a particularly economical means of implementing differences in TCoG-alignment. This, of course, in no way implies that it is the only means for so doing at the speaker’s disposal.

11 While TCoG ultimately aims to characterize both the timing and scaling of F0 events, most work to date has focused exclusively on timing contrasts.

12 In characterizing L*+H as a ‘delayed peak’, we do not intend to imply that the difference in High tone timing is the only thing that distinguishes the two pitch accents. In particular, while the phonetic details still strike us as equivocally demonstrated on this matter, there is nonetheless a commonly held sense (embodied, depending on one’s interpretation of it, in the starred tone notation), that while L+H* is a kind of High accent, L*+H is fundamentally a kind of Low. (See Ladd 2008’s repudiation of the analysis of Ladd 1983 for discussion.)
The first three participants were familiarized with the contours in a slightly different way: they were presented with dialogues of the kind illustrated above, representing either Incredulity or Uncertainty, which were read aloud by the experimenter through the beginning of the final line (i.e. the target phrase). At this point, they were played a recording of the monosyllabic target item embedded in a synthetic rise-fall-rise contour, the alignment and scaling parameters of which were designed to be ambiguous between L+H* and L*+H pitch accent timing. The idea was that context would cause the participants to categorize the contours according to the norms of their own usage, thereby producing L+H* or L*+H without being explicitly trained to do so. When it became clear that this approach was not as successful as had been hoped in orienting participants toward selection of particular phonological constructions, it was abandoned in favor of training with unambiguous versions of the contours in question.

Where multiple productions remained, which production to remove was decided by repetition number, meaning that, if a speaker’s second production of the target ‘lemonier’ with an L*+H pitch accent was excluded owing to a disfluency, then that subject’s second production of the same target with the contrasting L+H* pitch accent was removed as well.

It is normally desirable when dealing with perceptual properties of the F0 signal to use a scale such as semitones, rather than Hz. In the present context, this choice made little difference in our results.

House’s theory rests on the idea that certain properties of the signal (in particular, spectral stability or instability with a given region) are responsible for altering the robustness of tonal perception. His model, however, locates target windows for extracting average F0 values not directly via these properties, but rather using a temporally fixed window, during which he argues that the properties in question tend to be found.

Regarding the latter, it might be objected that, if it turns out that rises in these instances matter more than falls, this is simply because, under the standard analysis, the accents in question are fundamentally rising (i.e. comprised of an L+H sequence, as per Pierrehumbert 1980 and Ladd 2008, but note too the competing analysis of Gussenhoven 2004, et passim). This idea is contradicted, however, by evidence demonstrating some perceptual relevance of the shape of the fall in certain cases, such as that described below from D’Imperio’s (2000) work on Neapolitan Italian. It also fails to account for the phenomena observed by House, d’Alessandro, and Barnes and colleagues just described.

In addition, if it turns out in fact to be location with respect to the segmental string (e.g., proximity to some region of the accented vowel) that determines the perceptual weight accorded individual F0 samples, then we are also disadvantaging the model by limiting the TCoG window to the rise both for L+H* and L*+H pitch accents: While it is true that the rise for L+H* does tend to occur largely within the accented vowel in English, with the fall typically outside it, in the case of L*+H, both the rise and fall defining the high region tend to occur after the accented vowel, suggesting that sample weighting within the region of interest might differ between these two pitch accents. Again, since we are currently defining the window in terms of the F0 contour itself, rather than the segmental string, and since our goal here is to construct a model capable of categorizing pitch accents in an unsupervised fashion, we must apply the same procedure to all pitch accents, regardless of the location of the high region relative to the segmental string.

In our data set, the original TCoG systematically falls remarkably close to the automatically labelled F0 maximum (mean distance = 7 ms, SD = 19 ms). This makes sense, given that the peak, being the F0 maximum, is centered in the region of F0 samples contributing the highest weights to the calculation of TCoG, and hence attracting it in their direction.

d’Alessandro and colleagues, for example, while stating this as an explicit aim of their weighted time-average model of F0 target estimation, nonetheless achieve the greater influence exerted by higher F0 samples not through any direct system of weights applied, but rather in an indirect fashion, following from how weights are assigned to different portions of the glissando, depending on the direction of the movement, and the endpoint to be estimated.

It is important to note at this point that the application of this transformation to measured F0 values is not meant to represent anything like perceived F0 of the samples in question (to the extent that it is
meaningful to speak of such a thing during a tonal movement), but is instead solely for the purposes of increasing or decreasing the influence of certain samples in the derivation of TCoG. It is perhaps therefore better to think of the resulting value as something more like an attentional factor than like an $F_0$ value \textit{per se}.

22 Though we in fact ran this analysis multiple times, using all possible permutations of the TCoG model described above, we report here only the full model, with both the shortened analysis window, and the sigmatized $F_0$ weights implemented. While these refinements do not change our results dramatically, each did improve results sufficiently over a model without it that we judged it worthy of inclusion here.

23 Classification performance is derived by using the regression model to assign each utterance in the corpus a predicted probability of being, in this case, $L^*+H$. Utterances with a predicted probability over .5 are considered classified as $L^*+H$, while those under .5 are considered classified as $L+H^*$. Classification results are a relatively coarse way of looking at the performance of a model (insofar as they assess only whether predicted probabilities are over or under a given threshold, and not, for example, how unambiguously they fall to either side of it), and for this reason are not usually recommended as indicators of the goodness-of-fit of a model. On the other hand, classification results are intuitively comprehensible as indicators of a model’s success in a way that likelihood ratios and pseudo-$R^2$ measures are not, and for this reason, we will continue to report them here.

24 The TP-based model can in fact be improved slightly by adopting a different procedure for estimating the onset of the rise. This method involves locating not a local minimum in the $F_0$ curve, but rather an inflection point or ‘elbow’, marking something like the onset of the steeper portion of the rise. The method most frequently used for locating such elbows was developed by Mary Beckman and colleagues (Pierrehumbert and Beckman 1988; D’Imperio 2000; Frota 2002; Welby 2003; Welby and Loevenbruck 2006), and works by fitting two lines to the $F_0$ record in a region of interest, and minimizing the residuals from the least-squares fit between the $F_0$ values and the lines. Applied to our data, the TP-based model described above would see its percent correct classifications, for example, increase from 90.9% to 92.2%. There is an important caveat here, however. While elbows derived in this manner are usually employed simply as a more reliable means of estimating low TPs in the $F_0$ track, they are also crucially different from other TPs, in that their derivation, being a least-squares line-fitting procedure over a broad region of the contour, is in fact sensitive to many of the same aspects of global contour shape that affect TCoG. That is, unlike local maxima and minima, the $F_0$ elbow can be shifted, often a significant distance, by the shape of portions of the contour (e.g., the final portion of the rise before the peak) that lie a considerable distance away from the TP itself, and that would not ordinarily be considered relevant to the identification of a Low target. The use of $F_0$ elbows as Low TPs, though practically effective, in essence amounts to smuggling global contour shape in through the model’s back door. It is therefore not appropriate for inclusion in an assessment of the strengths and weaknesses of purely TP-based models.

25 It is interesting to note that the same procedure applied using the initial and untransformed TCoG-Rise models will select TCoG as a predictor first, but will then also select peak, with an extremely small, but statistically significant improvement to model fit. (Rise onset is rejected in all models.) The fact that the application of the sigmoid transformation to $F_0$ in the calculation of TCoG renders a significance of peak alignment where $p > .05$ (if only by a little: $p = .06$ here) suggests that this transformation has had the desired effect of sufficiently discounting the contribution of lower $F_0$s to the TCoG model.

26 Lack of error bars in the noise-free condition reflects the fact that this is just a single analysis of the base data set, identical to the one presented above in Experiment 1, rather than a mean taken over 30 distinct noisy relabellings of the base data set.

27 Or $H+L^*$, $(L+)^H^*$, and $L^*+H$ in the GToBI system for the analysis of German intonation (Grice, et al. 2005).

28 According to Niebuhr, who is not himself working in the GToBI framework, but rather uses the Kiel Intonation Model (Kohler 1991, 1995).

29 The contour shapes under comparison here are referred to by Niebuhr as slow rise/slow fall and fast rise/slow fall, respectively. Note, impressively, that this result obtained even when the accentual peak was
located within the vowel of the accented syllable, otherwise the touchstone of medial alignment for the relevant contours.

30 It should be noted, however, that Niebuhr and D’Imperio’s results also differed in important ways. For example, while fall duration in Neapolitan did influence listener categorization of accents as either L+H* or L*+H, fall duration did not have this effect in German for the accents bearing analogous labels. We believe the reason for this has to do with differences in the phonetic realization of L+H* and L*+H in the languages in question. In particular, both of these accents in Neapolitan, as described and modeled synthetically by D’Imperio, have peaks occurring far earlier with respect to the pitch-accented vowel than do their German representational analogues. Therefore, when D’Imperio constructs a ‘medial-to-late’ accent alignment continuum, the relevant steps of that continuum involve peaks ranging from 35% to 88% of the way through the (170 ms.) accented vowel. Even in the latest alignments, in other words, the peak is never delayed beyond the bounds of the pitch-accented vowel. As a consequence, the fall at issue here typically takes place within the vowel itself; throughout the continuum, the mean proportion of the fall taking place inside the pitch-accented vowel is 51% (maximum: 92%, minimum: 17%). In Niebuhr’s study of German, however, this picture was quite different: the largest proportion of the fall that occurs inside the pitch-accented syllable in any of the stimuli used in his medial-to-late alignment continuum appears to have been 33% (the earliest peak in this continuum is 40 ms from the end of the vowel, and the shortest fall is 120 ms). Indeed, since four of his seven continuum steps place the peak at or after the end of the pitch-accented vowel, in the majority of the cases no portion of the fall occurs with the vowel in question. Why this matters becomes clear in light of the discussion above in section 2.1.6.1 regarding asymmetrical F0 sample weighting as a function of location of samples relative to the segmental string: if it is the case that differences in fall duration matter more for categorization when that fall overlaps the accent vowel to a greater extent (as in Neapolitan), but don’t matter (or don’t matter as much) when the fall overlaps the accented vowel less, we might take this as evidence for the assignment of greater perceptual weight to F0 samples occurring within the accented vowel than to those falling outside it. This would be analogous to the result of Barnes and colleagues (forthcoming) involving sharp peaks and plateaux described above.

31 In fact, the situation is slightly more complicated than this, in ways that again likely relate to differences in sample weighting as a function of the alignment of the F0 contour with the segmental string: In the initial TCoG model presented above, all timepoints within the region of interest with a given F0 receive equal weight in the derivation of TCoG. This suggests that, for a plateau-shaped pitch accent perfectly symmetrical around its midpoint, TCoG should be located exactly halfway through the plateau. D’Imperio’s results, however, put the ‘point of equivalence’ between a sharp peak and a plateau somewhere much closer to the plateau’s end. Given the results described by Barnes and colleagues (forthcoming), however, it seems like that this pattern arises not because of any facts about the shape of plateaux per se, but rather as a function of how the specific plateaux tested in a given experiment are aligned with respect to the pitch-accented vowel.

32 Recall though that the ‘late’ L*+H of Neapolitan is nonetheless timed quite early in comparison with the analogous accents of languages like German or English: Most of the rise in Neapolitan L*+H typically takes place during the accented vowel, while in German or English, the rise of L*+H frequently begins only after the vowel’s offset.

References


Phonology 26. 43-74.


Barnes, Jonathan, Alejna Brugos, Nanette Veilleux, & Stefanie Shattuck-Hufnagel. 2011b. Modelling the perception of English F0 scaling in a segmental context. Experimental and Theoretical Advances in Prosody 2, Montreal, Canada.


Niebuhr, Oliver. 2007a. The Signalling of German Rising-Falling Intonation Categories – The Interplay of Synchronization, Shape, and Height. *Phonetica* 64, 2. 174-193.


Rossi, Mario. 1978. La perception des glissandos descendants dans les contours prosodique. *Phonetica* 35. 11-40.


Acknowledgements: We would like to thank Amalia Arvaniti, Jennifer Cole, Mariapaola D'Imperio, Carlos Gussenhoven, Oliver Niebuhr, and Yi Xu for helpful feedback on this work. We are likewise grateful to audiences at the first and second ETAP meetings in Ithaca and Montreal, Speech Prosody 2010 in Chicago, and the 17th ICPhS in Hong Kong, as well as to numerous anonymous reviewers. We also gratefully acknowledge all participants in our experimental work, and support from the National Science Foundation, grants #BCS-0643134, 0643054, 0643019, 1023853, 1023954, and 1023596.