

# Effects of pitch range variation on $f_0$ extrema in an imitation task

Laura C. Dilley<sup>a,\*</sup>, Meredith Brown<sup>b</sup>

<sup>a</sup>*Department of Communication Disorders, and Department of Psychology, Bowling Green State University, 247 Health Center,  
Bowling Green, OH 43403, USA*

<sup>b</sup>*Speech Communication Group, Massachusetts Institute of Technology, Building 36, Room 511, Cambridge, MA 02139, USA*

Received 13 September 2005; received in revised form 14 December 2006; accepted 2 January 2007

---

## Abstract

A central issue in speech intonation research concerns how fundamental frequency ( $f_0$ ) variation relates to phonological categories. The hypothesis was tested that pitch range variation which affects whether one syllable is higher or lower than another would elicit categorical shifts in  $f_0$  extremum timing in an imitation task. Participants heard synthetic versions of the phrase *Some lemonade* with rising-falling or falling-rising intonation and flat  $f_0$  contours across *le-* and *mo-*. The  $f_0$  levels of *le-* and *mo-* were varied such that for half the stimuli, *le-* had a higher  $f_0$  than *mo-*, while the reverse was true for the remainder. Participants produced  $f_0$  peaks and valleys on syllables that had flat  $f_0$  in stimuli; extremum types (peaks or valleys) and their temporal alignments varied categorically with the relative  $f_0$  levels of *le-* and *mo-* in the stimuli. The results are discussed in terms of theories of intonational phonology. It is shown that an account of these results under autosegmental-metrical theory (e.g., Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Ph.D. dissertation, MIT, Cambridge, MA) requires positing additional constraints in phonetic models of  $f_0$ . A revised version of the Pierrehumbert and Beckman [(1988). *Japanese tone structure*. Cambridge, MA: MIT Press] phonetic model is therefore proposed which assumes additional constraints on relative tone heights and strictly monotonic interpolation between tones.

© 2007 Elsevier Ltd. All rights reserved.

---

## 1. Introduction

A central issue in intonation research concerns how phonetic variation in fundamental frequency ( $f_0$ ) relates to phonological distinctions. At least two sources of  $f_0$  variation are typically distinguished in linguistic descriptions. While authors do not always agree on terminology, variation in  $f_0$  *contour shape* is generally taken to refer to differences in the pattern of rises and falls relative to syllables. Moreover, variation in *pitch range* generally refers to how high or low a given  $f_0$  curve is with respect to the overall limits that may be produced by a speaker. For example, Fig. 1 illustrates that pitch range can vary gradiently for a fixed, falling  $f_0$  contour shape (cf. Liberman & Pierrehumbert, 1984).

---

\*Corresponding author. Tel.: +1 419 372 7182; fax: +1 617 835 2441.  
E-mail address: [dilley@bgsnet.bgsu.edu](mailto:dilley@bgsnet.bgsu.edu) (L.C. Dilley).

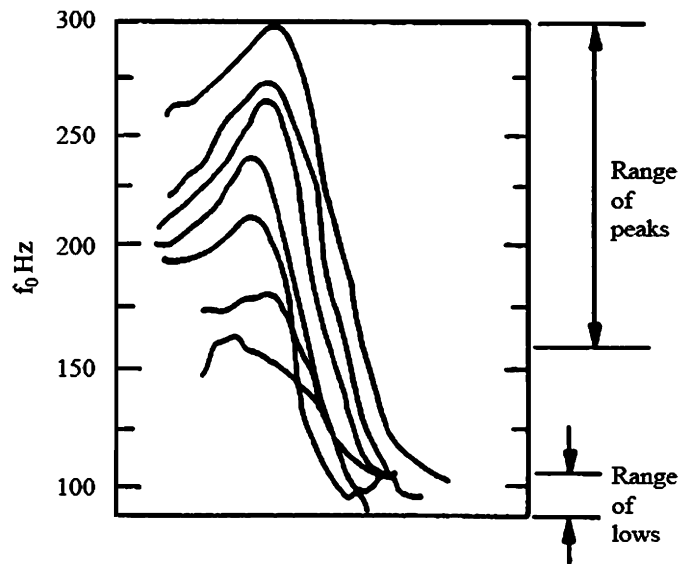


Fig. 1. Changes in pitch range under different degrees of emphasis. Reproduced from Liberman and Pierrehumbert (1984) with permission.

While linguistic descriptions have not necessarily treated variation in  $f_0$  contour shape and pitch range as independent, possible interactions between the two have rarely been discussed, let alone explored. The present paper explicitly addresses the issue of interactions between these two dimensions of variation, at both empirical and theoretical levels. Moreover, we seek to clarify the conditions under which pitch range variation may serve as a basis of phonological contrasts, as well as how any such contrasts are dealt with in theoretical terms, focusing here on English intonation.

Linguistic theories of intonation almost universally regard variation in  $f_0$  contour shape as phonologically contrastive. For example, the distinction between rises and falls is consistently treated as involving contrastive phonological categories, e.g., of questions vs. statements (e.g., Halliday, 1967; Pierrehumbert, 1980; ‘t Hart, Collier, & Cohen, 1990; see also examples and discussion in Gussenhoven, 1999). However, linguistic theories diverge when describing the relation between pitch range variation and phonological contrast. For example, pitch range variation is assumed to distinguish low rise and high rise contours by O’Connor and Arnold (1973). Moreover, Pierrehumbert (1980) has proposed that the lowering of heights of successive accents was due to a contrastive accentual type, while Ladd (1983, 1990) has maintained instead that downstep is a feature of vertical scale operating on otherwise identical accents. In general, however, pitch range variation has largely been treated as a gradient and non-categorical aspect of intonation (cf. Ladd, 1994, 1996).

In contrast to such generalizations, empirical studies suggest that pitch range variation may sometimes distinguish intonational meanings or categories. For example, Hirschberg and Ward (1992) showed that the size of a pitch excursion was the most significant of several phonetic factors determining incredulity vs. uncertainty interpretations of English rise-fall-rise contours (see also Ward & Hirschberg, 1985). Moreover, Bartels and Kingston (1994) suggest that tone height may be crucial in signaling the difference between L + H\* and H\* pitch accents in English. Ladd and Morton (1997) and Chen (2003) showed that the size of a rising-falling pitch excursion influenced both whether and how quickly English listeners interpreted synthetic speech stimuli as “normal” or “emphatic” accents. Finally, Vanrell Bosch (2006) used synthesized Majorcan Catalan utterances to show that varying the size of a pitch excursion resulted in an “s-shaped” identification function and a peak in a discrimination curve, suggesting that pitch range served as the phonetic basis of a phonological distinction (see also Bolinger, 1961; Gili Fivela, *in press*).

The first goal of the present paper is to clarify conditions under which pitch range variation may serve as a phonetic basis of English phonological contrast. In particular, we are interested in potential interactions between pitch range variation and  $f_0$  contour shape. Such interactions are possible under a class of theories in which the phonological representation is specified in terms of discrete tones or tone levels which are

time-aligned with segments or syllables (Beckman & Pierrehumbert, 1986; Liberman, 1975; Pierrehumbert, 1980; Pike, 1945). These theories share a common assumption, explicitly or implicitly, that  $f_0$  shapes arise from phonetic interpolation between phonologically specified discrete tones or levels. The aspects of tones that are assumed to be crucial for phonological representations include their positions in the pitch range (as e.g. high versus low) and/or their relative heights in sequence (Gussenhoven & Rietveld, 2000; Pierrehumbert, 1980). Such theories therefore contrast with those which take the view that rises and falls are phonological primitives, including the IPO approach (e.g., ‘t Hart et al., 1990) and the British school (e.g., Halliday, 1967).

For frameworks in which the discrete phonological entities are tones, under what conditions might pitch range variation be predicted to affect phonological representations? In considering such predictions, possible interactions between variation in pitch range and variation in  $f_0$  contour shape are often neglected. To understand such interactions, we will distinguish pitch range variations that alter the *relative heights* of the tones, i.e. whether one tone is higher or lower than another, from those that do not. First, we consider examples of pitch range variation which do *not* alter relative tone height (Fig. 2(a) and 2(b)). Here,  $T_1$  and  $T_2$  are two tones which are connected by phonetic interpolation; the vertical arrows indicate pitch range variability  $d$  in  $T_2$ . It can be observed that this pitch range variability does not affect the relative heights of the tones. For both contours in Fig. 2(a),  $T_1$  is lower than  $T_2$  and the expected overall  $f_0$  shape is rising, while for both contours in Fig. 2(b),  $T_1$  is higher than  $T_2$  and the expected overall  $f_0$  shape is falling. Given this type of variability, it turns out that discrete tone theories differ regarding whether pitch range variability will affect phonological analyses. On the one hand, the proposals of Pierrehumbert (1980), Beckman and Pierrehumbert (1986), and others, which have come to be known as autosegmental-metrical (AM) theory, treats pitch range variability of this sort as gradient variation, a view which has been termed the Free Gradient Variability Hypothesis by Ladd (1994, 1996). Thus the two contours in Fig. 2(a) would each consist of LH sequences, while the two contours in Fig. 2(b) would each consist of HL sequences.<sup>1</sup> In contrast, earlier theories by Pike (1945) and Liberman (1975) assume that there are four contrastive tone levels within a speaker’s pitch range; these theories would likely predict that pitch range variation of the type shown in Fig. 2(a) and (b) would be contrastive.

Next, consider the case in which pitch range variability exceeds certain limits such that it affects the relative heights of two tones, thereby generating distinct  $f_0$  shapes (Fig. 2(c)). Here, the change in relative heights of the tones associated with pitch range variability affects which  $f_0$  shapes are expected to arise through subsequent  $f_0$  interpolation. Thus (i) corresponds to a falling contour, while (ii) corresponds to a rising contour. Theories based on discrete tones are unanimous in their assumption that these two curves should have different phonological content, e.g. HL for (i) but LH for (ii) under AM theory. Thus, the first goal of the present paper is to clarify conditions under which pitch range variation may be contrastive, both when  $f_0$  shape is affected and when it is not.

While the theories of intonation discussed above view  $f_0$  shapes as arising from interpolation between discrete tonal targets, a great deal of recent attention within these theoretical frameworks has nevertheless been devoted to one aspect of  $f_0$  contour shape, namely  $f_0$  extremum alignment. Under one interpretation, an  $f_0$  extremum can be viewed as an aspect of  $f_0$  shape, which is localized to an individual segment or syllable. Extrema can be maxima, or peaks, or minima, or valleys; such extrema are often taken to be phonetic exponents of discrete tones (e.g., Arvaniti, Ladd, & Mennen, 1998; Dilley, Ladd, & Schepman, 2005; Ladd, Faulkner, Faulkner, & Schepman, 1999). A number of studies have now demonstrated that speakers show consistency in the timing or alignment of  $f_0$  extrema (e.g., Arvaniti et al., 1998; Atterer & Ladd, 2004; Dilley et al., 2005; Ladd et al., 1999; Ladd, Mennen, & Schepman, 2000; Lickley, Schepman, & Ladd, 2005; Xu, 1997, 1998). These alignment differences correspond to meaningful distinctions in many languages (e.g., D’Imperio, 2000; Grice, Ladd, & Arvaniti, 2000; Prieto, D’Imperio, & Gili-Fivela, 2005; Welby, 2003, 2006), and the timing of these extrema relative to segments is perceptually salient (e.g., Caspers & van Heuven, 1993; Dilley, 2005; House, 1990; Kohler, 1987; Niebuhr, 2003). Moreover, speakers imitate  $f_0$  extremum timing continua in a categorical way (Pierrehumbert & Steele, 1989; Redi, 2003). These data have generated

<sup>1</sup>More specifically, each rising contour might be analyzed as L\*H– (or L\*H–H%), while each falling contour might be analyzed as H\* L– (or H\* L–L%). In AM notation, “\*” denotes a tone which is associated with a stressed syllable, while “–” and “%” indicate a tone at the right edge of an intermediate or full intonational phrase, respectively.

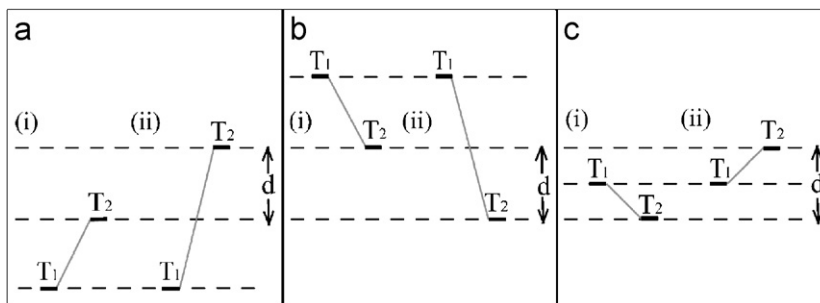


Fig. 2. Illustration of the effects of pitch range variability on the  $f_0$  shapes and relative heights of tones. See text.

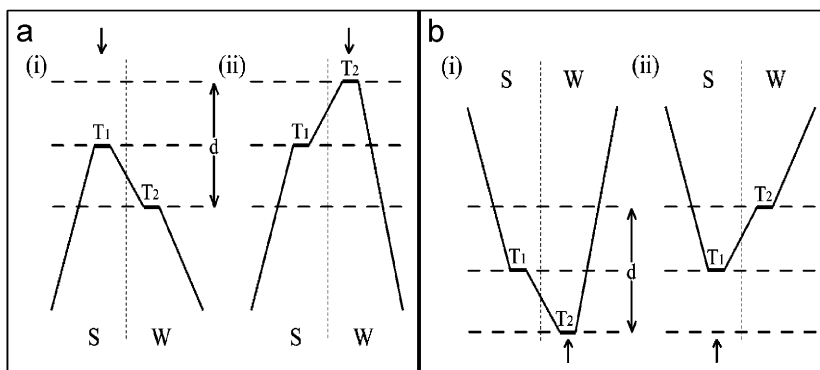


Fig. 3. Intonation patterns in which a change of relative height of two adjacent tones corresponds to a change in  $f_0$  extremum alignment, for theories in which the phonological entities are discrete tones. The dashed vertical lines indicate boundaries between strong (S) and weak (W) metrical positions.

considerable theoretical interest, since they can readily be accounted for in theories which assume that the underlying phonological entities are discrete targets synchronized with syllables, including both AM theory and the Parallel Encoding and Target Approximation (PENTA) model (Xu & Wang, 2001; Xu, 2005; Xu & Xu, 2005). PENTA assumes that  $f_0$  patterns arise as a result of the articulatory approach of either static or dynamic underlying targets. In contrast, such data cannot readily be accommodated under “contour-based” theories which assume that the phonological entities are rises and falls, e.g., the British school (Halliday, 1967) and the IPO approach (‘t Hart et al., 1990; see e.g., Ladd et al., 1999; Ladd, 2000 for discussions).

The second goal of the present paper is to explore a possible connection between two phonetic attributes which have previously been viewed as unrelated: *pitch range variation* and  *$f_0$  extremum alignment*. To see why  $f_0$  extremum alignment might potentially be related to pitch range variation, consider the contours in Fig. 3(a) and (b), which illustrate the same contours as in Fig. 2(c) but now preceded and/or followed by a rise or a fall. These figures show that a change in relative tone height affects the alignment of an  $f_0$  peak (Fig. 3(a)) or an  $f_0$  valley (Fig. 3(b)). In other words, *there is a close conceptual relationship* between pitch range variation, relative tone height, and  $f_0$  extremum alignment. Moreover, theoretical treatments predict that these factors should interact in producing phonological contrasts in a way which has not been previously explored in the literature.

The present paper explores this conceptual relationship by testing the hypothesis that perception of pitch range variation which affects the relative heights of two syllables underlies phonological contrastiveness demonstrated for  $f_0$  extremum timing. Such a hypothesis seems plausible, since relative syllable height is one of several phonetic attributes which covary with  $f_0$  extremum timing. For example, a syllable which exhibits a local  $f_0$  peak is expected to generally have a higher pitch than adjacent syllables (e.g., House, 1990). In addition, an  $f_0$  extremum is always redundantly preceded and followed by a dynamic pitch movement (i.e., a rise or fall). The relative importance of these co-occurring cues to phonological representations has not been

established, although there has been some work in this area (e.g., Grice & Savino, 1995; Niebuhr, 2003; Knight, 2003). Understanding the contributions of these covarying attributes is relevant not only for evaluating phonological theories, but also for modeling the interface between phonetics and phonology. Moreover, to the extent that a link might be demonstrated in perception and production between the relative pitches of syllables and  $f_0$  extremum timing, it is important for models of the phonetics-phonology interface to be able to provide an account of this link.

An imitation study was conducted to address these issues. Synthetic speech stimuli were constructed in which within-syllable cues to the presence and timing of an  $f_0$  extremum in the phrase *Some lemonade* were removed from each of two syllables, *le-* and *mo-*. Moreover, dynamic  $f_0$  cues preceding and following each syllable were removed by splicing in noise. The pitch of each syllable was then varied with respect to the speaker's pitch range and with respect to other syllables in sequence. We hypothesized that speakers would produce categorically distinct patterns of  $f_0$  extrema in response to relative pitch level cues across syllables, in order to test for the theoretically predicted interaction between pitch range variation, relative tone height, and  $f_0$  extremum alignment.

## 2. Methods

### 2.1. Participants

Participants were 13 students and staff at colleges in the Boston area (2 men, 11 women). All were self-reported native speakers of a general American English dialect. Moreover, all had self-reported normal hearing, and all were paid a nominal sum for participation.

### 2.2. Stimuli

Stimuli were created from tokens of the phrase *Some lemonade* produced by the first author, who is a native speaker of American English from the Midwest US. In the general American English dialect, the word *lemonade* may have main stress either on the first syllable or on the third syllable. For tokens of the phrase used in recordings, the main stress was produced on the initial syllable in *lemonade*. The mean  $f_0$  of the speaker for the selected tokens was approximately 225 Hz (range: 150–320 Hz). Recordings were made onto a DAT recorder with a 22.05 kHz sampling rate in a sound-attenuated chamber using a high-quality microphone; recorded utterances were subsequently transferred to a PC for synthetic manipulation.

The stimulus series shown in Fig. 4(a) and (c) were based on a token of *Some lemonade* produced with an overall rising-falling intonation pattern typical of a statement, while the stimulus series shown in Fig. 4(b) and (d) were based on a single utterance produced with an overall falling-rising intonation pattern typical of a question. The stimulus series were created by replacing the  $f_0$  contour across each of the two vowel nuclei in the critical two-syllable sequence *lemo-* with a level  $f_0$  contour, sometimes called a *plateau*, which was set to particular values as described below. The  $f_0$  contour across each syllable in synthetic speech materials was stylized using a sequence of straight lines and resynthesized using a pitch-synchronous overlap and add (PSOLA) algorithm implemented in Praat software (Boersma & Weenink, 2002; Moulines & Charpentier, 1990).

To create the stimulus series shown in Fig. 4(a) and (b), which we called the Roving-High and Roving-Low series, each of the syllables *le* and *mo-* was paired with a flat  $f_0$  pattern corresponding to one level in a 10-step series of  $f_0$  levels, where each of the levels in this series was separated by  $\frac{1}{2}$  semitone. Semitones were chosen as the basis for step sizes on the basis of evidence from Nolan (2003) that this scale best accounted for speaker behavior in an imitation task. Semitones were defined as a proportion such that for two frequencies  $f_1$  and  $f_2$  (in Hz) separated by  $n$  semitones,  $f_2 = 2^{n/12} f_1$ . The names “Roving-High” and “Roving-Low” are used here to reflect the fact that the  $f_0$  level on the first syllable was always different from trial to trial (hence it was “roving”). For each of the stimuli in the Roving-High and Roving-Low series, different  $f_0$  levels were selected for the first and second target syllables. For the Roving-High series, stimulus “1” was created by pairing the highest  $f_0$  level of the 10-step series on *le-* with the lowest level of the series on *mo-*, stimulus “2” was created by pairing the next highest  $f_0$  level on *le-* with the next lowest level on *mo-*, and so on, so that stimulus “10”

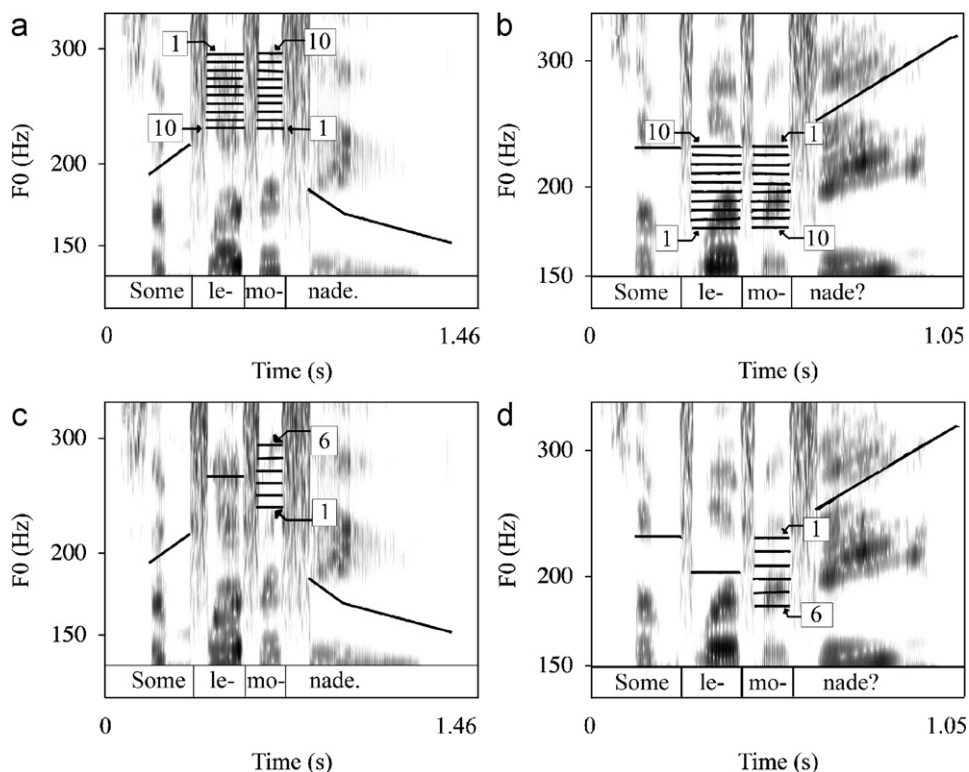


Fig. 4. Stimuli used in the present experiment: (a) Roving-High series, (b) Roving-Low series, (c) Fixed-High series, and (d) Fixed-Low series. The frequency scale for the spectrogram is 0 to 5000 Hz.

involved pairing the lowest  $f_0$  level on *le-* with the highest  $f_0$  level on *mo-*. In this way, stimuli “1” through “5” were predicted to give rise to early peaks, since for all five stimuli *le-* was higher than *mo-*, while stimuli “6” through “10” were predicted to give rise to late peaks, since for all five stimuli *mo-* was higher than *le-*. Similarly, for the Roving-Low series, stimulus “1” was created by pairing the lowest  $f_0$  level of the 10-step series on *le-* with the highest level of the series on *mo-*, stimulus “2” was created by pairing the next lowest  $f_0$  level on *le-* with the next highest level on *mo-*, and so on, so that stimulus “10” involved pairing the highest  $f_0$  level on *le-* with the lowest  $f_0$  level on *mo-*. In this way, stimuli “1” through “5” of the Roving-Low series were predicted to give rise to early valleys, since for all five stimuli *le-* was lower than *mo-*, while stimuli “6” through “10” were predicted to give rise to late valleys, since for all five stimuli *mo-* was lower than *le-*. Moreover, we predicted that participants would produce peaks in response to Roving-High stimuli and valleys in response to Roving-Low stimuli due to the fact that *le-* and *mo-* are high relative to neighboring syllables in the former case but low relative to neighboring syllables in the latter case. In other words, we predicted that not only the *alignment* of the  $f_0$  extremum, but also the *type* of extremum (peak or valley), would be influenced by the relative heights of *le-* and *mo-* with respect to each other and surrounding syllables for these two series.

To create the Fixed-High and Fixed-Low series in Fig. 4(c) and (d), *le-* was assigned a fixed, flat  $f_0$  level which was set at 262 Hz for the Fixed-High series and 202 Hz for the Fixed-Low series. Next, the second syllable, *mo-*, was then paired with one level in a 6-step series; each of the levels in this series was separated by  $\frac{3}{4}$  semitone. This manipulation ensured that the pitch range characteristics of the stressed syllable were identical within the series and that overall pitch range varied very little, in order to determine whether participants would respond to *le-* in different ways when only the  $f_0$  of *mo-* was varied. To create the Fixed-High series, stimulus “1” was created by pairing the lowest of the 6 levels with *mo-*, stimulus “2” was created by pairing the next lowest of the 6 levels with *mo-*, etc., such that stimulus “6” was created by pairing the highest of the 6

levels with *mo-*. In this way, stimuli “1” through “3” were predicted to give rise to early peaks, since for these three stimuli *le-* was higher than *mo-*, while stimuli “4” through “6” were predicted to give rise to late peaks, since for these stimuli *mo-* was higher than *le-*. Similarly, to create the Fixed-Low series, stimulus “1” was created by pairing the highest of the 6 levels with *mo-*, stimulus “2” was created by pairing the next highest of the 6 levels with *mo-*, etc., such that stimulus “6” was created by pairing the lowest of the 6 levels with *mo-*. In this way, stimuli “1” through “3” were predicted to give rise to early valleys, since for these three stimuli *le-* was lower than *mo-*, while stimuli “4” through “6” were predicted to give rise to late valleys, since for these stimuli *mo-* was lower than *le-*.

It was also necessary to remove  $f_0$  transition cues to the presence and timing of an  $f_0$  peak and valley to ensure that listeners could rely only on the relative  $f_0$  levels across syllables in the present experiment in producing any responses to stimuli. That is, even if an  $f_0$  peak on the vocalic nucleus of *le-* was replaced with a level  $f_0$ , we reasoned that listeners could still infer that an  $f_0$  peak had been present on this syllable based on residual transition information across syllable boundaries. To eliminate these additional  $f_0$  transition cues to the timing and presence of an  $f_0$  peak or valley,  $f_0$  information leading to and from the critical syllables *le-* and *mo-* was eliminated by splicing out the portions of each of the waveforms corresponding to [l], [m] and [n] at zero crossings and replacing these with white noise of identical duration. The white noise was generated using Praat’s sound-generation function with parameters  $\mu = 0$  and  $\sigma = 0.1$ , frequency range 0–11025 Hz, and a sampling rate of 22 050 Hz. The level for the noise was selected so that when spliced in, it sounded significantly louder than the initial fricative in *Some* but not so loud as to be uncomfortable when the remaining speech was presented at normal volume. The perceptual effect was that a consonant sound similar to [s] was inserted. A similar method of using noise to mask or replace  $f_0$  transitions was previously used by Xu and Xu (2003), who presented data suggesting that listeners likely interpret the  $f_0$  “jump” across the noisy interval as due to a smooth but inaudible laryngeal movement.

### 2.3. Procedure

An imitation task was employed, following a method similar to Pierrehumbert and Steele (1989) as well as Xu, Xu, and Sun (2004). Participants were told they would hear the phrase *Some lemonade*, as well as some noise. They were told to ignore the noise and to try to imitate the phrase that they heard as closely as possible in a comfortable pitch range. The latter instruction was included since some participants in a pilot study attempted to imitate the absolute pitch of the prompt stimuli. The participants’ attention was not otherwise drawn to the pitch of the stimuli.

Participants were seated in a sound-attenuated booth in front of a computer screen and a high-quality omnidirectional microphone was situated 8” from their lips to record their imitations. Auditory stimuli were presented directly from the PC’s hard drive using Winamp software.<sup>2</sup> Participants listened to each stimulus over high-fidelity headphones with the text of the target phrase simultaneously presented on the computer screen. The presentation rate was 1 stimulus every 4 s; the long inter-stimulus interval allowed sufficient time for participants to comfortably imitate the stimulus. The imitated utterances were digitized in real time using in-house, custom software by Mark Tiede (MARSHA v2.2 2002).<sup>3</sup> At the onset of each auditory presentation, the experimenter pressed a button on the computer keyboard to initiate a new recording buffer, the contents of which were automatically saved to the computer hard drive.

Over the course of the experiment, participants produced three imitations of each stimulus. Stimuli from the Roving-High and Fixed-High series were grouped together and randomized in a single test block (H), as were stimuli from the Roving-Low and Fixed-Low series (L). In total, this yielded six test blocks, which were presented in a fixed order: H, L, H, L, H, L.

Each test block was preceded by a set of practice trials consisting of stimuli drawn from the upcoming block. During practice trials, a few participants initially produced the target phrase with [s] in place of the noise (*Some sesosade*). These participants were corrected and reminded that the target phrase was *Some lemonade*. The experiment lasted about 25 min.

<sup>2</sup>Available at [www.winamp.com](http://www.winamp.com). Last viewed March 20, 2007.

<sup>3</sup>For information on the availability of this software, contact Mark Tiede at [tiede@haskins.yale.edu](mailto:tiede@haskins.yale.edu).

## 2.4. Analysis

The temporal positions of  $f_0$  peaks and valleys in participants' imitations were determined via visual inspection of  $f_0$  contour displays using Praat software, while the positions of syllable boundaries were determined via visual inspection of the spectrogram. The boundary between [m] and [l] in *Some le-* was taken as the location of an increase in amplitude across frequencies corresponding to the right edge of the nasal. When more than one position of discontinuity was observable, the position consistent with the greater amount of relatively low frequency energy was taken as the location of the boundary. The start and end of the [n] were marked separately.

Individual utterances were discarded from analysis on two grounds. First, the  $f_0$  contour across the target syllables was sometimes globally too flat to confidently determine the temporal location of a peak or valley. Second, participants sometimes failed to reproduce the final rising or falling intonation pattern, suggesting poor performance on the imitation task. If more than 50% of a participant's imitations were rejected on either of these two grounds, the participant's data for that series was withdrawn from analysis. This resulted in the removal of one participant from the Roving- and Fixed-High series and two participants from the Roving- and Fixed-Low series. Among the remaining participants, fewer than 5% of total data points were discarded.

The temporal position of peaks and valleys was then normalized with respect to segmental material using the formula in (1) in order to minimize any possible effects of speaking rate differences across participants. In the equation,  $T_N$  is the normalized  $f_0$  maximum or minimum time,  $d$  is the duration of *lemo-*,  $t$  is the start of [l] in *lemo-*, and  $t_0$  is the time associated with the  $f_0$  minimum or maximum. Because most participants consistently produced  $f_0$  extrema during or just after *lemo-*,  $T_N$  took values ranging from 0 to about 1. Dividing by the duration of the entire two-syllable target sequence allowed us to define a consistent normalization frame across all stimuli, regardless of whether  $f_0$  extrema were aligned with the first or the second target syllable in subjects' imitations:

$$T_N = \frac{t - t_0}{d}. \quad (1)$$

An initial inspection of the data suggested that a small number of participants were very poor imitators. In order to quantify this impression, a bivariate correlation analysis was conducted on  $T_N$ . This analysis yielded a correlation coefficient,  $r$ , for each pair of participants; a relatively high value  $r$  indicates that two participants produced similar patterns of data. For this analysis, we collapsed data from similar stimulus series: the Roving-High and Fixed-High series in one group and the Roving- and Fixed-Low series in another. Participants who were not reliably correlated at  $p < .05$  with half or more of the other participants were judged to be poor imitators, and their data were not included in the analysis. This resulted in the removal of two participants from the High series and one participant from the Low series, leaving ten participants in each of the four series.

Finally, to assess how accurately participants had reproduced the  $f_0$  differences across *le-* and *mo-* present in stimuli, the ratio of the average  $f_0$  level across *le-* to the average  $f_0$  level across *mo-* was determined for each utterance. A mean  $f_0$  ratio for each stimulus was then calculated, collapsing across participants.

## 3. Results

Our first finding concerns the types of  $f_0$  shapes produced by participants in this imitation experiment. Although the syllables *le-* and *mo-* had identical, locally flat  $f_0$  patterns in the stimuli, participants readily differentiated among these stimuli in their imitations. In particular, they produced  $f_0$  peaks in response to the Roving- and Fixed-High series and  $f_0$  valleys in response to the Roving- and Fixed-Low series. Fig. 5(a) and (b) show typical contours produced in response to stimuli from the two "high" series and the two "low" series, respectively. This effect was consistent across subjects with all responses to stimuli in Roving- and Fixed-High series being produced with peaks, as well as all responses to stimuli in Roving- and Fixed-Low series being produced with valleys.

Second, participants systematically varied the timing of peaks and valleys in response to the relative  $f_0$  levels of successive syllables in stimuli. Fig. 6 shows normalized  $f_0$  peak time,  $T_N$ , for each stimulus in the

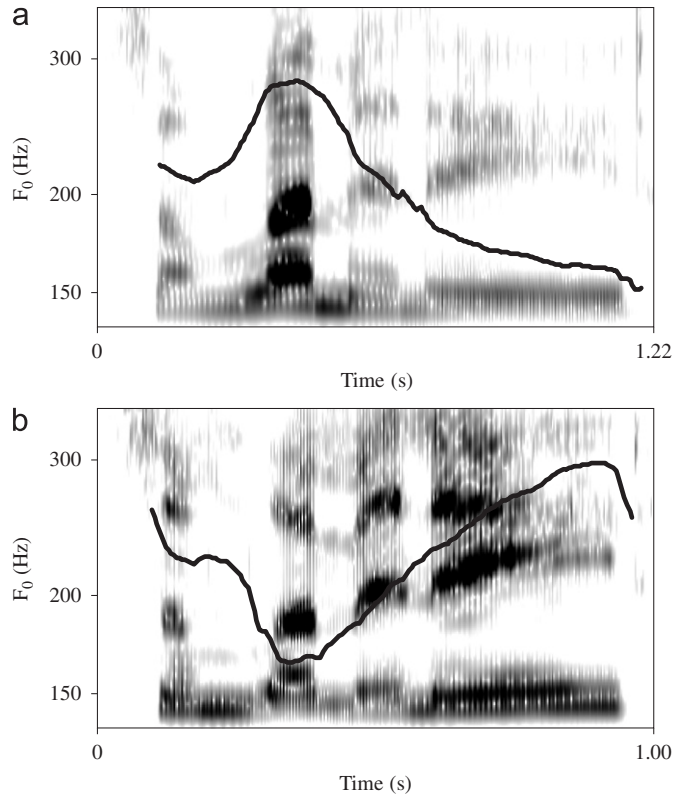


Fig. 5. Examples of typical imitations of stimuli in (a) the Roving-High and Fixed-High series and (b) the Roving-Low and Fixed-Low series. Shown are responses to the first stimulus in the Roving-High and Roving-Low series, respectively, for one subject. The frequency scale for the spectrogram is 0 to 5000 Hz.

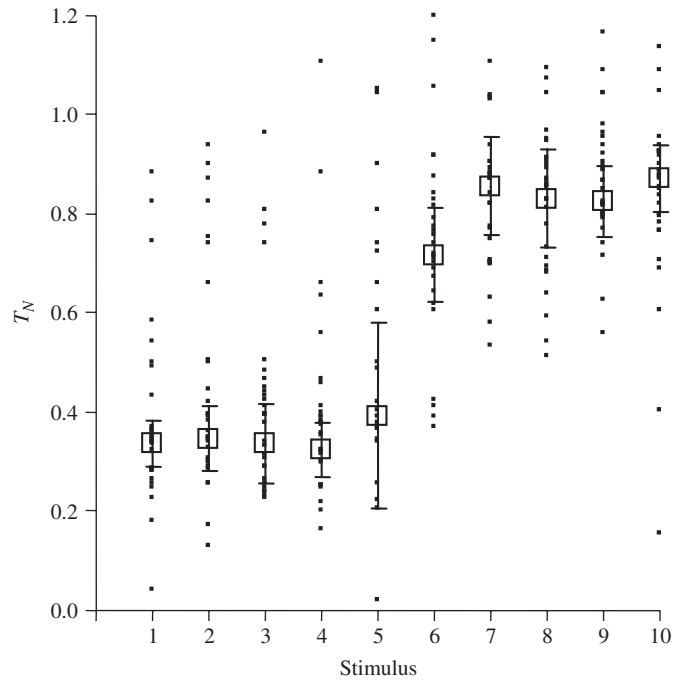


Fig. 6. Normalized extremum time,  $T_N$ , for the Roving-High series ( $n = 300$ ). Open squares show median values, while whiskers indicate the semi-interquartile range.

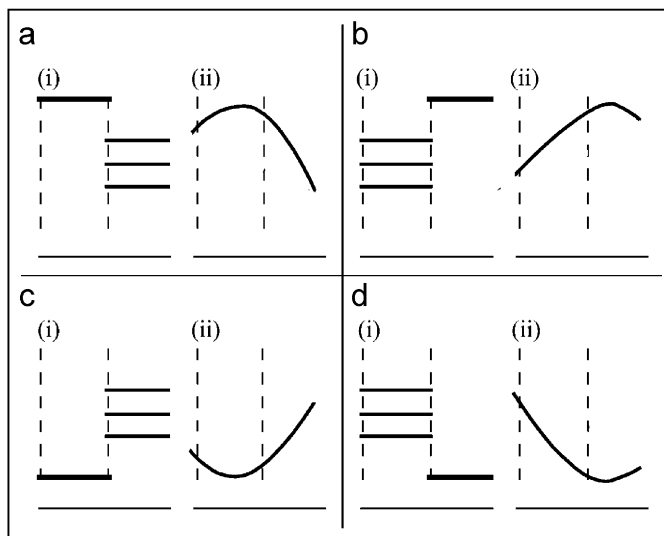


Fig. 7. Relationship between pitch levels in stimuli and  $f_0$  extremum type and alignment in imitations; vertical dashed lines indicate the start of *le-* and *mo-*, respectively. For the Roving-High and Fixed-High series, (a) shows that when *le-* was higher than *mo-* in stimuli (i), an early-timed  $f_0$  peak was observed in imitated versions (ii). Likewise, (b) shows that when *le-* was lower than *mo-* in stimuli (i), a late-timed  $f_0$  peak was observed in imitated versions (ii). For the Roving-Low and Fixed-Low series, (c) shows that when *le-* was lower than *mo-* in stimuli (i), an early-timed  $f_0$  valley was observed in imitations (ii). Finally, (d) shows that when *le-* was higher than *mo-* in stimuli (i), a late-timed  $f_0$  valley was observed in imitations (ii).

Roving-High series. Open squares indicate median values, while whiskers give the semi-interquartile range.<sup>4</sup> Here, participants produced  $f_0$  peak times that were relatively early in *lemo-* for stimuli 1–5 (as indicated by small  $T_N$ ) and  $f_0$  peak times which were relatively late in *lemo-* for stimuli 6–10 (as indicated by larger  $T_N$ ). This mirrors the relative  $f_0$  level relationships present in the Roving-High stimuli: for stimuli 1–5 the signal corresponding to *le-* had a higher  $f_0$  than that for *mo-*, while for stimuli 6–10 the opposite was true. The correspondence between  $f_0$  levels in the stimuli and peak times is shown in Fig. 7(a) and (b). A single-factor ANOVA confirmed the effect of stimulus on mean  $f_0$  peak time ( $F(1,9) = 33.664$ ,  $p < .0001$ ,  $MSE = .016$ ). Moreover, a planned *post hoc* comparison showed a significant difference in mean  $f_0$  peak time for stimuli 1–5 versus stimuli 6–10 in a paired-samples *t*-test ( $t(9) = -9.253$ ,  $p < .0001$ ).

Systematic variation of  $f_0$  extremum timing in relation to the relative  $f_0$  levels in stimuli was also obtained in responses to the Roving-Low series (Fig. 8). Here, participants produced  $f_0$  valley times which were relatively early in *lemo-* for stimuli 1–5 and relatively late in *lemo-* for stimuli 6–10. This pattern reflects the relative  $f_0$  level relationships in the Roving-Low stimuli: for stimuli 1–5 the signal corresponding to *le-* had a lower  $f_0$  than that for *mo-*, while for stimuli 6–10 the opposite was true, as depicted in Fig. 7(c) and (d). A single-factor ANOVA confirmed the effect of stimulus on mean  $f_0$  valley time ( $F(1,9) = 18.586$ ,  $p < .0001$ ,  $MSE = .022$ ). Moreover, a planned *post hoc* comparison showed a significant difference in mean  $f_0$  valley time for stimuli 1–5 versus stimuli 6–10 in a paired-samples *t*-test ( $t(9) = -6.919$ ,  $p < .0001$ ).

Comparable influences of relative  $f_0$  level on  $f_0$  peak and valley timing can be seen in responses to the Fixed-High and Fixed-Low series (Figs. 9 and 10). In Fig. 9,  $f_0$  peak times were relatively early in *lemo-* for stimuli 1–3 and relatively late in *lemo-* for stimuli 4–6; this mirrors the relative  $f_0$  level relationships present in the Fixed-High stimuli (Fig. 7(a) and (b)). The effect of Stimulus on mean  $f_0$  peak time is confirmed in a one-way ANOVA ( $F(1,9) = 18.586$ ,  $p < .0001$ ,  $MSE = .022$ ), while a planned *post hoc* comparison showed a significant difference in mean  $f_0$  peak time between stimuli 1–3 and 4–6 ( $t(9) = -9.388$ ,  $p < .0001$ ). Similarly, in Fig. 10,  $f_0$

<sup>4</sup>These measures were selected due to the non-normal distribution of the data. The semi-interquartile range (SIQR) is computed by taking one-half of the difference between the 75th percentile and the 25th percentile data points. Thus for a symmetric distribution, the interval stretching from one SIQR below the median to one SIQR above the median will contain  $\frac{1}{2}$  the scores.

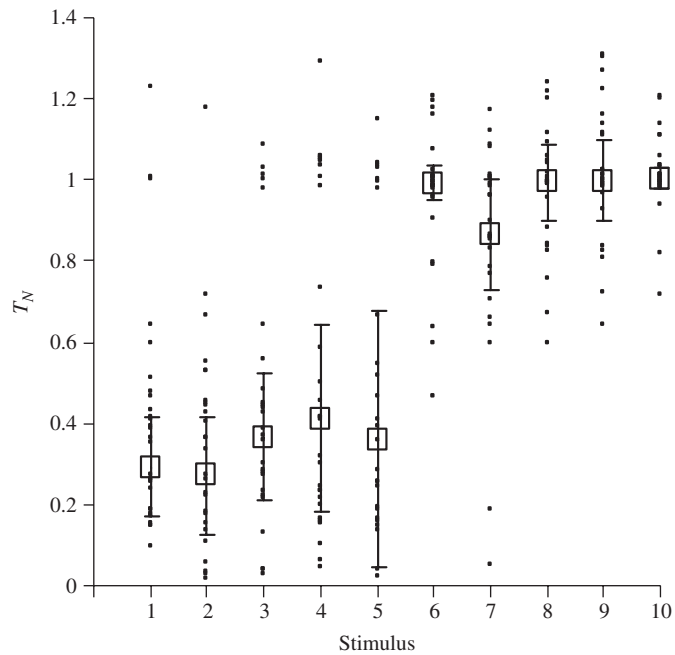


Fig. 8. Normalized extremum time,  $T_N$ , for the Roving-Low series ( $n = 237$ ). Open squares show median values, while whiskers indicate the semi-interquartile range.

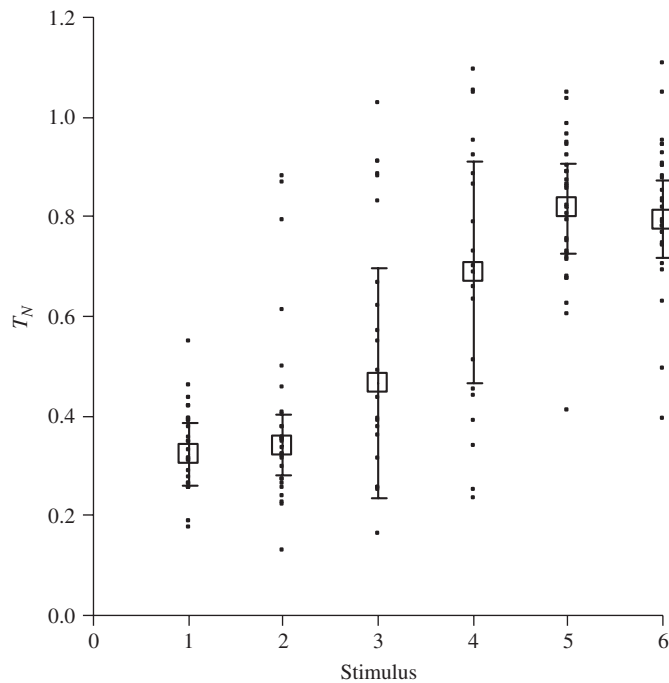


Fig. 9. Normalized extremum time,  $T_N$ , for the Fixed-High series ( $n = 180$ ). Open squares show median values, while whiskers indicate the semi-interquartile range.

valley times were relatively early in *lemo-* for stimuli 1–3 and relatively late in *lemo-* for stimuli 4–6; this pattern is consistent with the relative  $f_0$  level relationships present in the Fixed-Low stimuli (see Fig. 7(c) and (d)). Again, the effect of Stimulus on mean  $f_0$  valley time is confirmed in an ANOVA ( $F(1,5) = 29.214$ ,

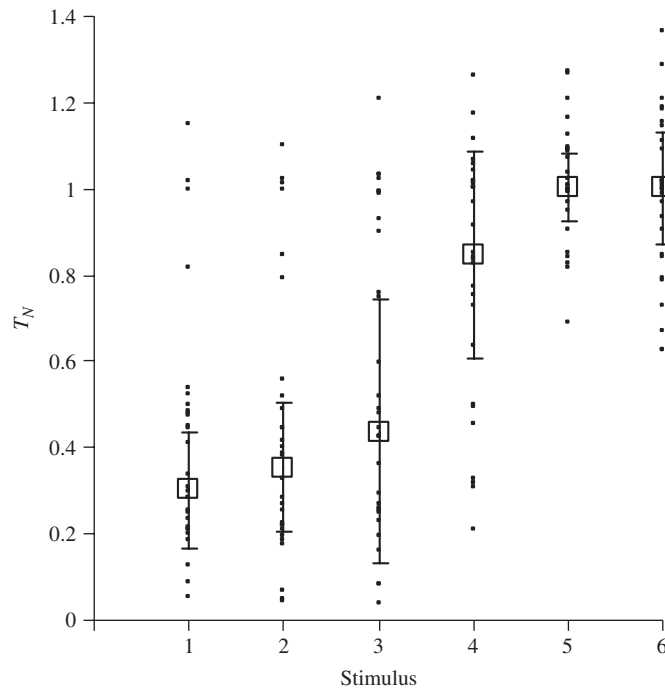


Fig. 10. Normalized extremum time,  $T_N$ , for the Fixed-Low series ( $n = 104$ ). Open squares show median values, while whiskers indicate the semi-interquartile range.

$p < .0001$ ,  $MSE = .025$ ); a planned *post hoc* comparison showed a significant difference in mean  $f_0$  valley time between stimuli 1–3 and 4–6 ( $t(9) = -7.575$ ,  $p < .0001$ ).

A third finding concerned consistency of timing of  $f_0$  valleys. Comparing the results from the two “High” series in Figs. 6 and 9 with the two “Low” series in Figs. 8 and 10, it is clear that  $f_0$  valley timing was comparable to that of  $f_0$  peak timing. The present study thus constitutes the first demonstration that participants are able to produce categorically distinct  $f_0$  valley timing in an imitation task. In a previous study, participants who imitated stimuli in which either  $f_0$  peaks or  $f_0$  valley times had been shifted along a continuum were only able to produce categorically distinct  $f_0$  peak times (Redi, 2003).

A fourth finding concerned how accurately participants reproduced the  $f_0$  intervals in the stimuli. Fig. 11(a) plots for the Roving-High and Roving-Low series the ratio of the average  $f_0$  level across *le-* and *mo-* in participants’ productions against the corresponding ratio in the stimuli. Participants showed some ability to reproduce the  $f_0$  interval in the stimuli in the range of stimuli 4–7 for both series, though there appeared to be somewhat better matching for the Roving-High than for the Roving-Low stimuli. Fig. 11(b) plots the ratio of the average  $f_0$  level across *le-* and *mo-* in participants’ productions for the Fixed-High and Fixed-Low series against the corresponding ratio in the stimuli. Here again, participants showed some ability to match the  $f_0$  interval for some stimuli, e.g. stimuli 2–5, for which the  $f_0$  levels on *le-* and *mo-* are relatively close.

#### 4. Discussion

This study investigated the phonetic and perceptual basis of phonological representations in intonation. An imitation study was conducted using stimuli in which dynamic  $f_0$  cues within and across key syllables were eliminated through replacement with a combination of flat  $f_0$  and white noise. This manipulation permitted an examination of the effects of different types of pitch range variation on participants’ imitations. We predicted that participants would imitate the level  $f_0$  patterns by producing  $f_0$  extrema (peaks and valleys), and that the presence, type, and timing of these extrema would be related in a categorical way to the relative  $f_0$  levels across the flat-pitch syllables.

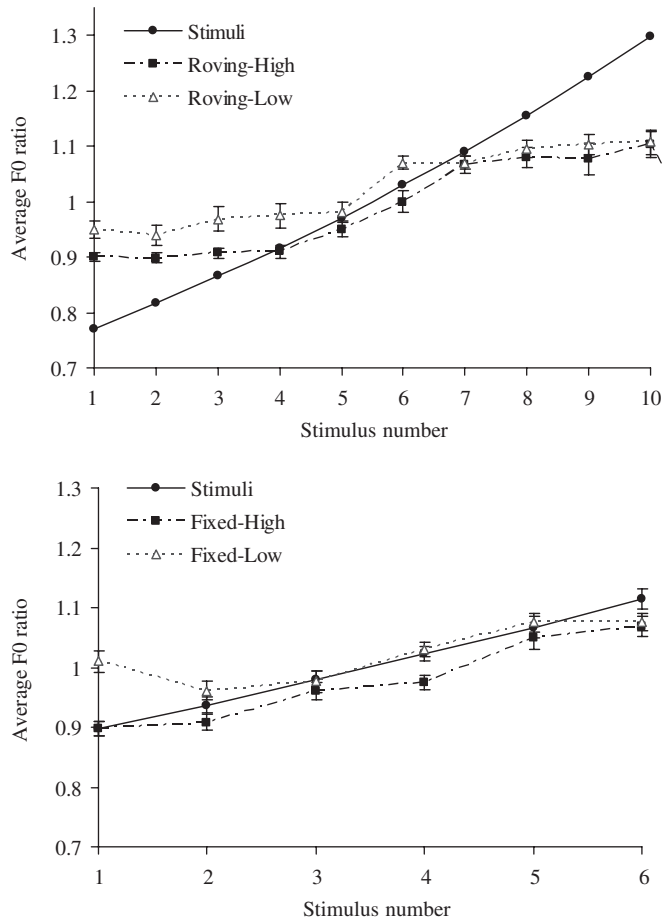


Fig. 11. Average ratio of  $f_0$  levels across *le-* and *mo-* for (a) the Roving-High series and Roving-Low series, and (b) the Fixed-High and Fixed-Low series.

The findings of this study clearly supported these predictions. Our first major result was that the relative pitch levels of the two target syllables in the stimuli determined whether subjects imitated a given flat-pitched target syllable *le-* or *mo-* by producing a single  $f_0$  peak, a single  $f_0$  valley, or no  $f_0$  extremum on that syllable. In particular, when a target syllable had a higher pitch relative to adjacent syllables, participants produced an  $f_0$  peak, but when a target syllable had a lower pitch relative to adjacent syllables, participants produced an  $f_0$  valley. These results indicate that subjects were encoding the syntagmatic relative pitch levels in the stimuli.

Our second main finding was that participants produced categorical temporal alignment for  $f_0$  peaks and valleys relative to segments, where this alignment varied consistently with the syntagmatic relative pitch levels of successive syllables in the stimuli. This is the first study to demonstrate effects of pitch range manipulations on categorical alignment in  $f_0$  extremum timing in a production task. Figs. 6, 8–10 show that timing of  $f_0$  extrema changed in a categorical way as syntagmatic relative  $f_0$  level varied within the stimulus series. For the Roving-High and Fixed-High series,  $f_0$  peak timing was early when the first target stimulus syllable had a higher  $f_0$  than the second target stimulus syllable, while  $f_0$  peak timing was late when the reverse was true. Conversely, for the Roving-Low and Fixed-Low series,  $f_0$  valley timing was early when the first target stimulus syllable had lower  $f_0$  than the second target stimulus syllable, while  $f_0$  valley timing was late when the reverse was true. We note that the categorical nature of the data was not dependent on whether  $f_0$  extremum times were normalized or not, nor on the choice of normalization method; researchers have variably preferred either normalized or non-normalized metrics (see e.g., Ladd et al., 1999; Silverman & Pierrehumbert, 1990; Xu, 1998). Finally, we note that categorical timing in imitation tasks has previously been demonstrated for  $f_0$

peaks only (Pierrehumbert & Steele, 1989; Redi, 2003). The present experiment is the first to demonstrate categorical timing for  $f_0$  valleys in speech.

A third main finding was that participants showed some limited ability to reproduce the gradient pitch range variation associated with the  $f_0$  interval between *le-* and *mo-* in the stimuli. Fig. 11(a) and (b) show that for stimuli in which the pitch interval between *le-* and *mo-* was small, participants reproduced the interval between the syllables with some accuracy. Outside of this range, however, participants showed little ability to reproduce the interval across syllables, instead producing approximately the same interval size in response to larger intervals.

One aspect of these findings was somewhat unexpected. While the relative  $f_0$  levels in stimuli were consistently related to the presence and timing of  $f_0$  peaks and valleys in participants' imitations, responses to the Fixed-High and Fixed-Low stimulus series appeared to show a less abrupt category boundary than responses to the Roving-High and Roving-Low stimulus series. This is likely due to the fact that the minimum difference in  $f_0$  levels across the target syllables *le-* and *mo-* was smaller for the Fixed-High and Fixed-Low series than for the Roving-High and Roving-Low series ( $\frac{3}{8}$  s.t. vs.  $\frac{1}{2}$  s.t., or a difference of about 6 vs. 8 Hz). In response to stimuli in the Fixed-High and Fixed-Low series for which the  $f_0$  levels across target syllables were very close (i.e., stimuli 3 and 4 from each series), a couple of participants produced the target syllables so that they had audibly the same pitch, rather than producing the syllables so that one syllable clearly had a higher or lower pitch than the other syllable. This resulted in producing an  $f_0$  peak or valley which had a normalized time that was intermediate between the other cases, resulting in the appearance of a less abrupt category boundary. (See Knight (2003) for a discussion of similar phenomena.) The pattern of results is still very similar to that for the Roving-High and Roving-Low cases, but qualitatively displays an "S-shaped" curve which is nevertheless typical in experiments involving categorical perception and production.

The present study is the first to document categorical effects in  $f_0$  extremum timing in an imitation task in response to pitch range manipulations. Previous studies which have demonstrated categorical effects in  $f_0$  extremum timing have relied on stimuli in which  $f_0$  extremum timing *per se* was varied along a continuum (Dilley, 2005; Pierrehumbert & Steele, 1989; Redi, 2003). Categorical effects in perception and/or production of  $f_0$  peaks and valleys have long been taken as evidence of distinct phonological categories in intonation (cf. Bruce, 1977; Gussenhoven, 2004; Kohler, 1987; Ladd, 1996; Pierrehumbert & Steele, 1989). Moreover, categorical effects in timing of  $f_0$  extrema in an imitation task are considered the main indicators of phonological contrastiveness (Gussenhoven, 1999, 2004). The categorical behavior in  $f_0$  extremum alignment can thus be interpreted as evidence that the stimuli in this study cued a phonological distinction.

Can these results be explained without assuming that participants encoded the syntagmatic relative  $f_0$  levels across successive syllables? For example, might participants in this experiment have simply imitated the absolute  $f_0$  of each syllable in succession? The answer is "no", since participants reproduced stimuli in their own pitch ranges instead of imitating absolute  $f_0$ . This observation suggests that participants were indeed imitating the syntagmatic relative pitch levels of syllables in the stimuli.

Our finding that relative pitch height induced categorical effects in  $f_0$  extremum alignment is important because it suggests that a phonetic cue which covaries with extremum alignment—namely, the relative pitch height of syllables—may be responsible for certain categorical effects which have previously been associated with peak or valley timing *per se*. Usually,  $f_0$  extremum timing covaries with other phonetic variables, including dynamic pitch information and relative pitch information across syllables. In this experiment, we removed these covarying phonetic variables—including both  $f_0$  extrema and dynamic pitch information—from the stimuli in order to examine the resulting effects on of relative pitch target realization. This experiment demonstrates that  $f_0$  extrema are not necessary in order to cue phonological contrast; rather, relative pitch level is sufficient. Future work will be necessary to determine the relative significance of alignment cues vs. relative pitch cues, since such findings have implications for theories of how listeners recover phonologically relevant intonational information from speech.

The present results extend earlier work on how speakers imitate level pitches in several ways. Xu and Sun (2002) investigated how quickly speakers could produce changes in pitch, utilizing a task in which speakers imitated a very rapidly alternating sequence of level high and level low pitches on a sustained schwa vowel or syllable sequence. Under these conditions, Xu and Sun showed that speakers imitated alternating high and low flat pitches by producing  $f_0$  peaks and valleys, respectively. However, Xu and Sun's experiment did not

investigate a possible link between pitch range variation and  $f_0$  extremum alignment. Moreover, the experiment left unresolved whether speakers would produce  $f_0$  peaks and valleys using more natural speech materials and at a more normal rate of speech. Finally, their experiment included only two pitch levels, which were always globally highest or lowest, raising the question e.g., of whether participants would produce peaks and valleys on all level-pitched syllables which were high in the pitch range or low in the pitch range, or merely the highest and lowest. Our experiment showed that when the globally high or low pitch occurred on the first syllable in the target sequence, speakers produced early-timed  $f_0$  peaks or valleys; conversely, when the globally high or low pitch occurred on the second syllable in the target sequence, speakers produced late-timed  $f_0$  peaks or valleys. Moreover, the present results suggest that speakers imitate  $f_0$  levels as  $f_0$  peaks and valleys even at normal speech rates and while producing real speech phrases.

The results showed that participants could reproduce the pitch range of stimuli to a limited extent, that is, when the interval is small. This finding is consistent with the idea that this pitch range matching was due to some degree of encoding of within-category gradation. Such gradation is predicted under the Free Gradient Variability Hypothesis (Ladd, 1994, 1996). Thus some evidence supports the notion that pitch range variation is gradient when the contour shape is expected to remain the same. However, the present results also showed that outside of a narrow range, there is no evidence of quantitative matching of the pitch range. At least two possible explanations may account for this behavior, which we cannot distinguish between at the present time. The first is that speakers were treating the level variation as gradient for all stimuli and encoding the pitch distances between the levels, but that production accuracy was diminished for larger intervals. This account is predicted by the PENTA model, according to which biomechanical sluggishness of the vocal cords would be expected to limit dynamic accuracy in reproducing larger pitch intervals. (See discussion in Section 4.1 below.)

Another explanation for poor accuracy in reproducing large pitch intervals is that speakers do not perceive syllables as having pitch targets when the intervals are large. This would be expected if large, monotonic changes in pitch across successive syllables were perceived as interpolation. Under this explanation, speakers show limited accuracy in reproducing large pitch intervals because they cannot encode the pitch distances on syllables that occur in the middle of large monotonic pitch changes.

Yet another possibility suggested by a reviewer was that speakers were perceiving pitch intervals in a categorical way analogous to lexical tones such that they reproduced two pitch level categories, leading to poor interval matching accuracy in imitation. The roughly sigmoidal character of Fig. 11(a) could be taken to suggest that stimuli 1–3 and 8–10 correspond to distinct “within-category” regions, while stimuli 4–7 correspond to an “across-category” region. However, we disagree with this interpretation of the data, since in order for the analogy with categorical perception to hold, the so-called “across-category” region should be characterized by categorical perception. Thus, participants should have exhibited poor ability to perceive gradient variation in this region. However, this is precisely the region in which we find the best quantitative matching of gradient variation, in contrast to the categorical perception interpretation. However, we must also consider whether the mere *appearance* of accuracy arises as a result of averaging across stimuli which were perceived categorically. If so, then speakers should have reproduced stimuli 4–7 close to the “category boundary” with more variable pitch intervals, but “within category” stimuli 1–3 and 8–10 would be produced with less variable pitch intervals. To test this possibility, we compared the variability as measured by standard errors for stimuli 4–7 vs. stimuli 1–3 and 8–10, collapsing across the Roving-High and Roving-Low stimulus series. The difference between these groups was not significant in an independent-samples *t*-test ( $t(18) = -0.904, p = .378$ ), suggesting that an interpretation of these data in terms of categorical perception of distinct tone levels is not viable.

Having described the general implications of these results for understanding the phonetic basis of phonological distinctions, we consider the specific implications of these results for PENTA and AM theories.

#### 4.1. Interpretation of the data within the PENTA framework

The PENTA (Parallel Encoding and Target Approximation) model proposes that each syllable is associated with an articulatory pitch target which is temporally coordinated with the onset and offset of a syllable (Xu, 2005; Xu & Xu, 2005; Xu, 1997, 1998, 1999, 2001; Xu & Sun, 2002; Xu & Wang, 2001). Pitch targets can be static or dynamic; there are three types of static targets ([High], [Mid], and [Low]) and two types of dynamic

targets ([Rising] and [Falling]). According to the theory, targets are asymptotically approached in a linear fashion starting from the onset of a syllable, so that there will be a variable temporal delay between the syllable onset and when speakers are able to achieve the underlying target. The exact locations of  $f_0$  peaks and valleys will depend on underlying target type, as well as tonal context, syllable duration and articulatory strength (Xu, 2002). The specific  $f_0$  values produced for a given contour are assumed to be derived from multiple parallel communicative functions, such as sentence type (e.g., statement vs. question), focus, and lexical status (when applicable). Each communicative function is associated with a separate encoding function which specifies the parameters of the Target Approximation model.

The results overall are quite compatible with PENTA. Under the assumption that every syllable had a target, speakers produced contours in which every syllable was connected by a linear interpolation function. In particular, a single high or low  $f_0$  extremum is expected if speakers are producing linear interpolations between targets on each syllable, as assumed under PENTA. However, specifically which targets appear on individual syllables in the stimuli is not clearly predictable from PENTA, since the phonetic and perceptual basis of distinctive targets has not been completely clarified in this model. Descriptions of targets have emphasized distinctive patterns of peak alignment, although Xu and colleagues have implied that individual targets are also distinguished through a combination of  $f_0$  characteristics, including peak and valley alignment, global pitch range, and relative height. The present results may be interpreted as helping to clarify the nature of targets under PENTA by demonstrating a role for relative pitch in distinguishing tonal targets from one another (e.g., as [High] vs. [Rising]).

One aspect of the alignment data does not support PENTA's predictions. PENTA assumes that targets are articulatorily implemented with respect to syllable offsets and onsets, predicting that timing of  $f_0$  extrema should be more consistent when gauged with respect to a single syllable (*le-* or *mo-*, depending on the stimulus) than with respect to a two-syllable sequence *lemo-*. To test this prediction, we determined the timing of  $f_0$  extrema with respect to a single syllable (i.e., with respect to *le-* for stimuli 1–5 in the Roving-High and -Low series and stimuli 1–3 in the Fixed-High and -Low series, and with respect to *mo-* for the remaining stimuli). We then compared the accuracy of timing under the PENTA method to that of the two-syllable timing normalization method reported in the Results. To estimate relative accuracy, we calculated a coefficient of variation for peak and valley time for both the one-syllable normalization method preferred under PENTA and the two-syllable method employed here, where coefficient of variation is defined as the standard deviation divided by the mean across participants. The coefficient of variation was greater for the one-syllable PENTA method than for the two-syllable normalization method: 0.392 vs. 0.317, respectively. This difference was significant in a paired-samples *t*-test ( $t(31) = -3.353$ ,  $p < .0001$ ). Thus the prediction of PENTA that timing will be less variable when normalized with respect to a single syllable is not supported by the present data. One possible explanation for this finding is that speakers of English are more variable in their timing of  $f_0$  peaks and valleys than speakers of Mandarin, the language on which many PENTA model predictions have been developed.

#### 4.2. Interpretation of the data within the AM framework

The present results carry several implications for AM theory. First, these results do not support a strong interpretation of AM theory, according to which tones are phonetically equated with  $f_0$  extrema. Such a strong interpretation has occasionally been implied in the literature in that  $f_0$  peaks and valleys have been treated as the direct phonetic exponents of underlying H and L tones (see e.g., Dilley et al., 2005; Ladd et al., 1999). In the present experiment, cues to the timing and presence of  $f_0$  peaks and valleys were removed by replacing them with flat pitch, while splicing in noise at syllable boundaries to remove redundant dynamic  $f_0$  cues. This manipulation left only relative pitch cues intact across syllables. If  $f_0$  peaks and valleys were necessary for conveying contrastive intonational patterns, speakers would not have responded with categorical timing of  $f_0$  peaks and valleys in their imitations. Instead, the categorical behavior produced by the subjects suggests that relative pitch is at least as important as  $f_0$  extrema in signaling phonological distinctions. That  $f_0$  extrema are not necessary components of intonational patterns is consistent with work showing that pitch patterns sometimes show plateaus rather than peaks (D'Imperio, 2000; Knight, 2003). Because listeners perceive differences in the slopes of  $f_0$  contours across syllables (e.g., Niebuhr, 2003; Knight, 2003), it seems likely that

listeners would perceive the level pitches across target syllables in these stimuli as distinct from syllables with actual  $f_0$  peaks and valleys. Thus the  $f_0$  “plateaus” in these stimuli were not likely to have been perceived as perceptually identical to stimuli with  $f_0$  extrema, but merely to have had the same phonological representation.

We can also consider the implications of our finding that individuals showed some accuracy in reproducing pitch range variation over a limited range. This finding neither confirms nor disconfirms the Free Gradient Variability Hypothesis, i.e., the proposal that pitch range variation that does not affect  $f_0$  shape does not affect phonological category membership (Ladd, 1994, 1996). Given that a number of possible explanations may account for the limited accuracy shown here in reproducing pitch intervals, it will be necessary to investigate the validity of the FGVH in future studies.

How can the descriptive phonological framework of AM theory accommodate the present results? This theory proposes that  $f_0$  contours are comprised of a sequence of H and L tones which are either prominence-lending pitch accents or phrase-related accents and boundary tones. Pitch accents may consist of a single tone (H\* or L\*), or they may be bitonal. “Starred” tones, which are written with an asterisk, e.g., L\*, associate with stressed syllables, and thus predict temporal coordination between the  $f_0$  exponent(s) of the tone (e.g., an  $f_0$  valley) and that syllable. Moreover, the unstarred tones within bitonal pitch accents (e.g., L+ in bitonal L+H\*) lead or trail starred tones and fall on metrically weak positions. Thus it can generally be stated that differences in types of  $f_0$  extrema are explained in AM theory in terms of differences in tonal types (i.e., H vs. L), while differences in alignment are explained in terms of differences in tones’ status as starred vs. unstarred.

Based on standard descriptions of AM tonal sequences in the literature, including those associated with the ToBI (Tones and Break Indices) transcription system (e.g., Beckman & Pierrehumbert, 1986; Beckman & Ayers Elam, 1997), the categorical differences in types and alignments of  $f_0$  extrema demonstrated in this experiment can be accommodated in terms of the distinctive patterns of pitch accents and phrasal tones shown in Table 1. Differences in the alignment of  $f_0$  peaks can be described in terms of differences in the type of H tone: as H\* on *le-* for the early peak vs. H+ on *mo-* in a bitonal H+!H\* (i.e., H+L\*) anchored to *-nade* for the late peak. Likewise, differences in the alignment of  $f_0$  valleys can be described in terms of differences in the type of L tone: as L\* on *le-* for the early valley vs. L+ on *mo-* in a bitonal L+H\* anchored to *-nade* for the late valley. These ToBI analyses were independently provided for the stimuli by an experienced ToBI labeler naïve to the purposes of the present experiment. Note that these accent analyses indicate that the location of the strongest syllable was perceived to vary, depending on the location of the  $f_0$  maximum or minimum. This is consistent with findings from a perception study using American English listeners which showed that the timing of an  $f_0$  peak or valley influenced which syllable was perceived as strongest in polysyllabic words with ambiguous relative stress, e.g. *MILLionaire* vs. *millioNAIRE* (Dilley, 2005; Shattuck-Hufnagel et al., 2004). Recall that in general American English, the main stress in the word *lemonade* can be on the first or the third syllable. Thus AM theory can account for the present results in terms of distinctive phonological categories of representation.

Table 1  
Description of distinctive patterns of  $f_0$  extrema and timing produced by participants in terms of AM pitch accent and boundary tone sequences for each of the four stimulus series in the experiment

Stimulus series	$f_0$ extremum type	Extremum timing	AM phonological description
Roving-High and Fixed-High	Peak	Early	Some <i>le- mo- nade.</i> <div style="margin-left: 100px;"> <math>\begin{array}{c}   \\ H^* \end{array}</math> </div>
Roving-High and Fixed-High	Peak	Late	Some <i>le- mo- nade.</i> <div style="margin-left: 100px;"> <math>\begin{array}{c}   \\ L-L\% \end{array}</math> </div>
Roving-Low and Fixed-Low	Valley	Early	Some <i>le- mo- nade?</i> <div style="margin-left: 100px;"> <math>\begin{array}{c}   \\ H+!H^* \end{array}</math> </div>
Roving-Low and Fixed-Low	Valley	Late	Some <i>le- mo- nade?</i> <div style="margin-left: 100px;"> <math>\begin{array}{c}   \\ L^* \end{array}</math> </div>
			<div style="margin-left: 100px;"> <math>\begin{array}{c}   \\ H-H\% \end{array}</math> </div>
			<div style="margin-left: 100px;"> <math>\begin{array}{c}   \\ L+H^* \end{array}</math> </div>
			<div style="margin-left: 100px;"> <math>\begin{array}{c}   \\ H-H\% \end{array}</math> </div>

Finally, we note that the category of H\* proposed in [Pierrehumbert \(1980\)](#) is sometimes assumed to be realized with “peak delay”, in which the  $f_0$  peak generated for the high tone is realized temporally after the end of the stressed syllable ([Beckman & Ayers Elam, 1997](#)). Such a temporal “lag” in the timing of the peak relative to the stressed syllable is assumed to correspond to within-category variation for H\*, although [Dilley \(2005\)](#) has presented evidence against this interpretation. Nevertheless, the fact that H\* is assumed to sometimes be produced with peak delay raises the question of whether stimuli 6–10 and 4–6 for the Roving-High and Fixed-High series, respectively, might have been interpreted in terms of H\* accents. We reject this possibility on the grounds that the categorical nature of the timing of  $f_0$  extrema in responses to the first vs. second half of each stimulus series are incompatible with an interpretation solely in terms of a single category of H\*.

Having discussed the phonological interpretation of these results under AM theory, we turn in the following section to the issue of how these categories are assumed to be mapped to  $f_0$  values. In particular, we show that previous AM phonetic models provide insufficient constraints in mapping tones to  $f_0$  values, leading to problems in accounting both for the present data, as well as previous findings of consistency in  $f_0$  extremum timing. Finally, in Section 4.4 we present a revised version of the [Pierrehumbert and Beckman \(1988\)](#) phonetic model which appropriately constrains phonetic mapping functions, thereby providing an account of these results.

#### 4.3. *The treatment of syntagmatic relative tone height under AM phonetic models*

To fully evaluate how well AM theory accounts for the relationship between syntagmatic relative pitch and distinctive tonal categories in this experiment, these phonological analyses should not be considered in isolation, but instead in the context of the entire theory. A critical aspect of the AM theoretical approach is a phonetic component which scales tones in the pitch range and interpolates between them. The phonetic module is an integral component of AM theory which must be evaluated alongside the phonology:

...the division of labor between the phonology and the phonetics is an empirical question, one which can only be decided by constructing complete models in which the role of both in describing the sound structure is made explicit. ([Pierrehumbert & Beckman, 1988, p. 4](#))

The phonetic component under AM theory provides a stated algorithm for mapping discrete phonological representations to gradient phonetic variables. Previous research has established that the presence and timing of  $f_0$  extrema is important for phonological representations across many languages (see [Ladd, 1996, 2000](#) for reviews). In this section we inquire how the phonetic component of AM theory deals with the mapping of  $f_0$  values to phonological representations, and vice versa. In particular, we examine the extent to which AM phonetic models support the phonological analyses given in [Table 1](#), with respect to both the types of  $f_0$  extrema produced by participants (peaks vs. valleys), as well as their alignments. What we will see is that, generally speaking, previous AM phonetic models provide insufficient syntagmatic constraints on mapping tones to  $f_0$  values, leading to problems in accounting for our experimental results in terms of specific phonological categories.

Descriptions of tonal categories in AM theory commonly assume restrictions on tone scaling, particularly with respect to the relative heights of tones. For example:

...a pitch accent can impose a particular relationship between the  $f_0$  on the accented syllable and the immediately preceding or following  $f_0$  value, independent of the existence of other accents. In [[Bolinger's, 1958](#)] theory, all pitch accents are like this, and they are accordingly described in terms of  $f_0$  changes. In our theory, the bitonal accents have this property and there are also two single tones which do not. ([Pierrehumbert, 1980, p. 14](#))

In addition to bitonal pitch accents, restrictions on the relative heights of tones are commonly assumed to be in place in sequences like H\* L– and L\* H–. That is, H\* is assumed to never fall below adjacent L–, while L\* is assumed to never rise above adjacent H–. Assumptions of this sort are widespread in the literature and can be found in [Pierrehumbert \(1980\)](#), [Beckman & Pierrehumbert \(1986\)](#), and others. To what extent are these widely assumed syntagmatic restrictions on relative pitch height actually instantiated in AM phonetic models?

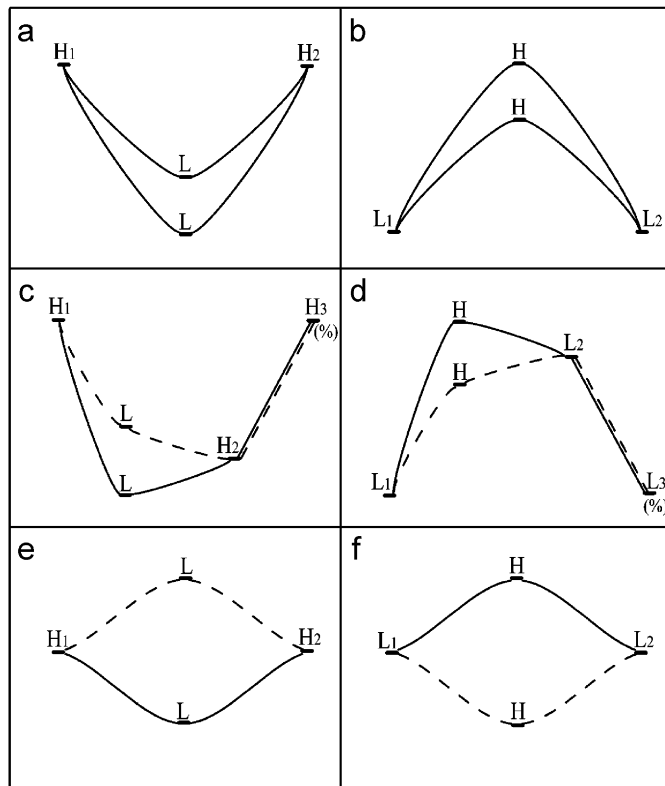


Fig. 12. Effects of type of relative height constraints on possible contours arising from tonal sequences plus interpolation. Dashed lines indicate contours which are degenerate outputs.

Before we address the issue of what restrictions are in place in phonetic models of tone mapping under AM theory, we first examine how *insufficient* relative height restrictions might affect  $f_0$  outputs, particularly with respect to  $f_0$  extremum type and alignment. Consider sequences of adjacent low-high and high-low tones in HLH(H) and LHL(L) contexts. Fig. 12(a)–(f) show several cases in which pitch range variability affects the relative height of a given tone with respect to one or both adjacent tones, or neither tone. First, Fig. 12(a) and (b) illustrate the expected outputs for  $H_1LH_2$  and  $L_1HL_2$ , respectively. In these contours, the relative heights of L and H tones are constrained with respect to both the leftward and the rightward high and low tones, respectively. Next, Fig. 12(c) illustrates pitch range variability in L for the context  $H_1LH_2H_3$ . Here, L does not rise above leftward  $H_1$  under this pitch range variability, but the alignment of an  $f_0$  valley is nevertheless affected, depending on whether the L falls below or above  $H_2$ . If L falls below  $H_2$ , then the expected alignment is generated, with the L corresponding to an  $f_0$  valley. However, if L falls above  $H_2$ , then an unexpected alignment pattern is generated, with  $H_2$  corresponding to the location of the  $f_0$  valley. This degenerate output is drawn with a dashed line. Similarly, Fig. 12(d) illustrates pitch range variability in H for the context  $L_1HL_2L_3$ . Here, H does not fall below leftward  $L_1$  under this pitch range variability, but the alignment of an  $f_0$  peak is nevertheless affected, depending on whether the H falls above or below  $L_2$ . If H is above  $L_2$ , then the expected alignment is generated, with the H corresponding to an  $f_0$  peak. However, if H falls below  $L_2$ , then an unexpected alignment pattern is generated, with  $L_2$  corresponding to the location of the  $f_0$  peak. This degenerate output is drawn with a dashed line. Finally, Fig. 12(e) and (f) and illustrate the effects of pitch range variability in L and H tones, respectively, when their relative heights are constrained with respect to neither the leftward nor the rightward tone. Fig. 12(e) illustrates that the contour expected to arise from  $H_1LH_2$  can correspond either to the expected fall-rise or an unexpected rise-fall, generating either an  $f_0$  peak or an  $f_0$  valley in the vicinity of L. Similarly, Fig. 12(f) illustrates that the

contour expected to arise from  $L_1HL_2$  can correspond either to the expected rise-fall or an unexpected fall-rise, generating either an  $f_0$  peak or a degenerate  $f_0$  valley in the vicinity of H.

These figures crucially illustrate that in order for the type and alignment of an  $f_0$  peak or valley to be predictable from the tonal sequence, the syntagmatic relative heights of every pair of tones must be restricted. This suggests that an account of the present data and previous results demonstrating consistent  $f_0$  extremum alignment require that sufficient restrictions on syntagmatic relative tone height be in place to generate the requisite phonological shapes. Do phonetic models within AM theory support intuition about the  $f_0$  extremum types and alignments which should result from phonological analyses in terms of high and low tones, while at the same time preventing degenerate phonetic outputs?

The Appendix addresses these questions through detailed analyses of the two prominent phonetic models of tone scaling under AM theory, namely [Pierrehumbert \(1980\)](#) and [Pierrehumbert and Beckman \(1988\)](#), which are hereafter referred to as P80 and PB88, respectively. We focus on these models since they alone provide accounts of the scaling of both H and L tones.<sup>5</sup> In the Appendix we have provided analyses of constraints on relative tone height for six critical tonal sequences in AM theory: bitonal pitch accents  $L+H^*$ ,  $L^*+H$ ,  $H+L^*$  and  $H^*+L$ , plus the single tone plus phrase accent sequences  $H^*L-$  and  $L^*H-$ . The analyses in the Appendix extend work recently reported by [Dilley \(2005, 2006\)](#).

For each AM tonal sequence, the critical question is whether there are sufficient syntagmatic restrictions on scaling adjacent tones, that is, whether L is prevented from rising above adjacent H, and whether H is prevented from falling below adjacent L. For the P80 model, the evaluations in the Appendix show that the syntagmatic relative heights of adjacent H and L tones are constrained in a way that prevents degenerate outputs only for *two sequences out of the six examined*:  $H^*L-$  and  $H+L^*$ . For  $H+L^*$ , the analyses also revealed a theoretical gap, in that the  $f_0$  contour for this accent is undefined in all positions except phrase-initial position. The analyses also showed that for  $L+H^*$  and  $L^*+H$ , the relative heights of the two tones are unconstrained for all phrasal positions. Finally, theoretical gaps and/or exceptional treatment were shown to lead to undefined  $f_0$  contours for  $L^*H-$  and  $H^*+L$  in all phrasal positions. For the PB88, the analyses revealed that there are no restrictions on the relative heights of adjacent H and L tones. Thus L is permitted to be higher than adjacent H in all positions both within and across phrases.

The analyses in the Appendix represent the first quantitative demonstration of the problematic effects of insufficient constraints on relative tone height in phonetic models. Previously, [Ladd \(1990, 1993, 1996\)](#) has alluded to problems with AM phonetic models, stating that “unconstrained gradient variability of pitch range parameters [is] the most serious empirical weakness of a great many quantitatively explicit models of  $f_0$ ” (1990, p. 37). However, no details regarding the consequences of unconstrained variability have been provided.

The analyses in the Appendix therefore demonstrate that in order for AM theory to adequately account for consistency in  $f_0$  peak and valley alignment, modifications to existing proposals will be necessary. In particular, phonetic and/or phonological mechanisms for restricting syntagmatic heights of adjacent tones must be in place so that consistency in the types and alignments of  $f_0$  extrema in phonetic outputs are predictable from phonological inputs. The need for such syntagmatic restrictions is demonstrated in [Fig. 12\(c\)–\(f\)](#). Without such additional restrictions, not only can the present results not be accounted for, but neither can a large body of phonetic findings from the past two decades demonstrating consistency in  $f_0$  peaks and valleys (see [Ladd, 1996, 2000](#) for reviews). This is because unconstrained syntagmatic relations between adjacent tones in phonetic models lead to unpredictable alignments and types of  $f_0$  extrema.

One possible objection that might be raised to the analyses provided in the Appendix is that the phonological definitions of H and L tones in AM theory automatically constrain the phonetic values of these tones so that, e.g., L does not rise above adjacent H. Close inspection of autosegmental theory ([Goldsmith, 1976](#); [Williams, 1971/6](#)) indicates that relative tone heights are not formally constrained in the phonology in this way. Autosegmental theory claims that tones are exactly like segments: they are autonomous segments in their own right (i.e., they are *autosegmental*). This assumption entails two separable claims. The first claim was that tones are autonomous from segments but are temporally coordinated with them. The second claim was that tonal features are exactly like segmental features. When autosegmental theory was first put forward,

<sup>5</sup>The phonetic model of [Lieberman and Pierrehumbert \(1984\)](#) is not considered, since it focused only on the scaling of H tones and did not give a quantitative account of L tone scaling.

segmental features were assumed to be strictly paradigmatic (i.e., non-relational) following *Sound Patterns of English* (Chomsky & Halle, 1968). Thus, autosegmental theory entailed what might be called a *strong paradigmatic* view, in which there is no formal syntagmatic interpretation of L and H which predicts that e.g., L is lower than H (see Snider & van der Hulst (1992) for discussion). In contrast to these strong paradigmatic formalisms, Pierrehumbert (1980) put forward a *weak paradigmatic* descriptive tonal system, one which clearly intended a role for syntagmatic tonal features (as the quote from p. 14 above illustrates). However, this syntagmatic role has nowhere been instantiated in formal phonological proposals within standard AM theory. Ladd (1990, 1993) has proposed syntagmatic phonetic restrictions on tone scaling for pitch range parameters at the phrasal level, but has not addressed the issue of syntagmatic scaling of adjacent H and L tones.

The finding of the present experiment that the relative pitch levels in the stimuli influence the realization of tonal contours in English in a categorical way clearly indicates a role for syntagmatic tonal features in the *phonological* representation, not just the phonetics. Thus, the present experimental results provide evidence against a strong paradigmatic view of English tonal features, instead supporting a weak paradigmatic interpretation of tones in which there is a role for syntagmatic tonal features. Such a weak paradigmatic assumption is clearly what was intended in the original proposals of Pierrehumbert (1980); however, formal phonological treatments of syntagmatic restrictions on tones have not yet found their way into mainstream AM theory. Proposals for incorporating syntagmatic features into tonal phonology which build on autosegmental tenets have recently been put forward by Snider (1999) and Dilley (2005, 2006, to appear).

#### 4.4. A revised version of the PB88 phonetic model

To address the problem of insufficient syntagmatic constraints on adjacent tones in previous AM models demonstrated in the Appendix, a revised version of the PB88 phonetic model is proposed here which constrains the relative heights of adjacent H and L tones within a phrase. Under the original model of PB88, each tone was assigned prominence value, which determined the tone's  $f_0$  level relative to reference  $f_0$  values. A number of studies, however, suggest that there is a complex, indirect mapping between perceived prominence and  $f_0$  level within a speaker's pitch range (Gussenhoven & Rietveld, 1988, 2000; Gussenhoven, Repp, Rietveld, Rump, & Terken, 1997; Ladd, Verhoeven, & Jacobs, 1994). Therefore, in our revised model we assume instead that the mediating gradient phonetic parameter is *pitch*, denoted  $P(T)$  for "pitch of tone  $T$ ", where we invoke the notions of perceptual and psychophysical pitch in our definition (cf. Moore, 2003). We assume that prominence influences  $P(T)$ , but does not directly determine it. One advantage of assuming that pitch is the mediating parameter is to bring AM phonetic models closer to the perceptual experience of the listener.

The value of  $P(T)$  determines the position of a high or low tone with respect to two abstract  $f_0$  reference lines: a high tone line,  $h$ , and a low reference line,  $r$ , where  $h > r$ . H and L tones are scaled with opposite polarity relative to  $h$  and  $r$ . The equations stating the relationship between  $f_0$  and pitch for H and L tones are given in (2) and (3). For H tones,  $P(H) = 1$  implies that  $f_0(H) = h$ , while  $P(H) = 0$  implies that  $f_0(H) = r$ .

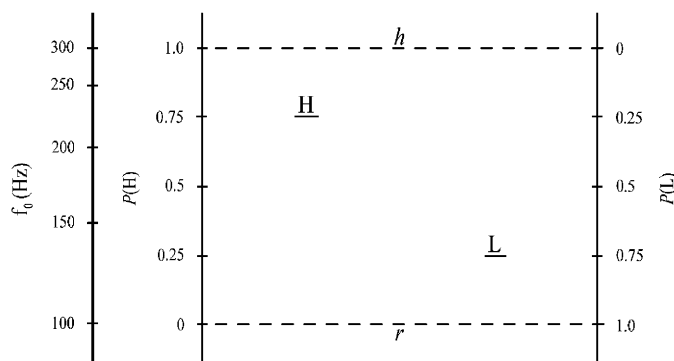


Fig. 13. Scaling of H and L relative to reference lines  $h$  and  $r$  in the original and revised phonetic models of Pierrehumbert and Beckman (1988). Here,  $h$  and  $r$  are set arbitrarily to 300 Hz and 100 Hz, respectively. See text.

For L tones, in contrast,  $P(L) = 1$  implies that  $f_0(L) = r$ , while  $P(L) = 0$  implies that  $f_0(L) = h$  (see Fig. 13). Note that the nonlinear relationship between  $f_0$  and pitch implies that the higher the absolute  $f_0$ , the larger the  $f_0$  change must be in order to generate an equivalent pitch change at lower absolute values of  $f_0$ . Finally, we define parameters  $h_{\max}$  and  $r_{\min}$  corresponding to the highest and lowest modal  $f_0$  values, respectively, which can be produced by a given speaker. This addresses a minor problem with the model of PB88, which imposed no restrictions on absolute  $f_0$  values; thus, more  $f_0$  values were predicted to be possible for any one speaker than could be produced physiologically. The equations giving the relationship between H and L tones,  $f_0$ , and pitch for a revised version of the PB88 model are stated below:

$$f_0(H) = [P(H)][h - r] + r, \quad (2)$$

$$f_0(L) = [1 - P(L)][h - r] + r. \quad (3)$$

Moreover, we propose that all phonetic interpolation functions between adjacent tones are monotonic, in contrast to the original proposals of Pierrehumbert (1980). This is consistent with recent experimental findings by Ladd and Schepman (2003), as well as Dilley (2005). In the original work of Pierrehumbert (1980), exceptional treatment was proposed for sequences of two high tones, which were assumed to be connected by a “sagging”, or nonmonotonic, interpolation function; in contrast, all other tone pairs were assumed to be connected by a monotonic interpolation function. The phonological distinction between contours involving two high accents separated by a mere “sagging” transition vs. a high-low-high tonal sequence as in  $H^* L + H^*$  was assumed to be made through two sorts of phonetic differences. First,  $f_0$  minima arising from a “sagging transition” were assumed to occur temporally mid-way between the two accents, while those associated with a low tone in  $L + H^*$  were assumed to be aligned just before a high accent. Second, the two contours were assumed to be distinguished through a difference in pitch range: the pitch range from the  $f_0$  minimum to a following high maximum was assumed to be larger when the minimum corresponded to a low tone, and relatively smaller otherwise. The assumption of “sagging transitions” was predicated largely on theory-internal assumptions about the nature of phonological “triggers” for lowering of successive accents, or “downstep” (Ladd, 2000); many of these assumptions have since been rescinded (Beckman & Pierrehumbert, 1986).

Two kinds of recent evidence converge to support a more unified AM treatment in which  $f_0$  minima uniformly correspond to low tones, and transitions between tones are strictly monotonic. First, Ladd and Schepman (2003) showed that English speakers apparently always produce  $f_0$  minima between two high accents so as to be aligned just before the second high accent. This supports a description in which the  $f_0$  minimum is consistently a low tone. Second, Dilley (2005) recently conducted an imitation study which suggested that pitch range differences in the height of relatively low-pitched unaccented syllables preceding a relatively higher accent are not interpreted as categorical, but rather as gradient differences. In the experiment, listeners heard the phrase *Some oregano* with a relatively low pitch across the initial two syllables *Some or-*, followed by a peak on *reg-* and a final fall. In the general American English dialect spoken by the participants, the single main stress in *oregano* is on the second syllable. The pitch range of the initial two syllables was varied along a continuum, while the remaining contour was held fixed. The stimuli thus ranged from canonical  $H^*$  at one end of the continuum to canonical  $L + H^*$  at the other end of the continuum. In the imitation task, participants reproduced the gradient present in the continuum, showing no evidence of categorical imitation of pitch range. The experiment failed to support the original claim of Pierrehumbert (1980) that pitch range is the basis of a categorical distinction between  $H^*$  and  $L + H^*$  accents. Taken together, Dilley’s experiment and that of Ladd and Schepman support a model in which (1)  $f_0$  minima are uniformly treated as low tones and (2) all interpolation functions are monotonic, as proposed here.

The critical aspect of our model that leads to restrictions on the relative heights of adjacent H and L tones is the stipulation that *for all L tones adjacent to H tones within an intermediate or full intonational phrase*,  $P(H) > 1 - P(L)$ . This formulation has the advantage of permitting phonological inputs to generate predictable phonetic outputs in terms of  $f_0$  extremum type and alignment. It therefore accounts for the results demonstrated in this experiment by stipulating syntagmatic restrictions on the relative heights of adjacent tones in a way that permits predictable  $f_0$  extremum types and alignments to be generated from phonological representations.

At least two lines of argument, however, suggest that a different solution will ultimately be needed to the problem of accounting for the demonstrated relationship between syntagmatic relative tone height and  $f_0$  extremum type and timing. First, examples can be found across languages in which patterns described by a HL or LH tone sequence do not necessarily correspond to a falling or rising  $f_0$  contour. For example, in French  $L_1H_1L_2H_2$  contours, the  $f_0$  often falls from  $H_1$  to  $L_2$ , but it can also be level or occasionally even rise (Welby, 2003). Similar variability in relative scaling has been reported for the French  $H^*$  and  $H_i$  (Rolland & Lævenbruck, 2002; Welby, 2003, 2006). Second, the present experiment suggests that syntagmatic restrictions are part of the phonological component of the grammar, since listeners interpreted whether a tone was higher or lower than another tone as contrastive. Thus it will ultimately be necessary to develop a theoretical account in which syntagmatic properties are part of the phonological representation of tones themselves. In this regard, the proposals of Snider (1999) and Dilley (2005, 2006, to appear), which formally instantiate syntagmatic representations as part of phonology, might be considered starting points for continuing to build on the rich insights already presented through AM phonological approaches.

## 5. Summary and conclusion

This imitation experiment demonstrated categorical effects in  $f_0$  peak and valley timing in response to stimuli in which (1) cues on individual syllables to  $f_0$  extrema were removed and (2) the relative  $f_0$  levels across syllable pairs were manipulated. Such categorical effects in production are considered the main indicators for phonological contrastiveness (Gussenhoven, 2004; Pierrehumbert & Steele, 1989). These results suggest that  $f_0$  extremum timing cues within syllables are not necessary for cueing phonological contrast, and that the relative  $f_0$  levels across successive syllables are instead sufficient. These results can be incorporated under the general theoretic assumptions of both PENTA and AM theories. However, inspection of phonetic models accompanying AM theory reveals that these models include insufficient syntagmatic constraints on scaling relative heights of adjacent tones, leading to problems in quantitatively accounting for the observed relationship between phonological representations and  $f_0$  extremum type and alignment. To address these issues, a revised version of the phonetic model of Pierrehumbert and Beckman (1988) is proposed. Given the evidence that syntagmatic tonal relations are part of phonology, however, a phonological treatment of syntagmatic relations will ultimately be needed.

## Acknowledgements

We owe tremendous thanks to Jonathan Harrington, Pilar Prieto and two anonymous reviewers for helpful comments which greatly improved the paper. Thanks also to Stefanie Shattuck-Hufnagel and Ken Stevens for many useful discussions and insights, and Aniruddh Patel for additional helpful feedback on an earlier draft. Moreover, we acknowledge Mara Breen for providing ToBI transcriptions of the stimuli. This research was supported by an NIH Training Grant awarded to the Speech and Hearing Bioscience and Technology Program, Harvard-MIT Division of Health Sciences and Technology, and by the MIT Undergraduate Research Opportunities Program.

## Appendix. Evaluation of treatment of relative tone height in AM phonetic models

This Appendix presents analyses of restrictions on syntagmatic relative tone heights for bitonal  $L+H^*$ ,  $L^*+H$ ,  $H+L^*$  and  $H^*+L$  pitch accents, as well as for  $H^*L-$  and  $L^*H-$  in the phonetic models of Pierrehumbert (1980) and Pierrehumbert and Beckman (1988). Appropriate restrictions on relative tone height are required both in order to account for data on consistent  $f_0$  alignment, such as that presented in this paper, as well as to prevent degenerate qualitative  $f_0$  outputs (see Section 4.3 for discussion).

### A.1. The Pierrehumbert (1980) phonetic model

We first examine the extent of syntagmatic restrictions on scaling relative tone heights in the phonetic model of Pierrehumbert (1980), or P80. In that model, the  $f_0$  value of each tone is computed relative to the  $f_0$  values of

previous tones. The  $f_0$  value of the first tone in a phrase is a free choice (p. 144). Each subsequent tone is assigned a value of a parameter termed *prominence*, which determines that tone's  $f_0$  according to a set of context-dependent phonetic rules. Ladd (1993, 1996) has noted that prominence as used in the P80 model has little to do with perceptual prominence. However, the P80 model clearly intends prominence to be a scalar variable with some acoustic or perceptual interpretation, one which was amenable to mathematical interpretation and quantitative formulation. In the following, we are crucially concerned with what restrictions are placed on prominence values in the P80 model, since such values ultimately determine the syntagmatic relative  $f_0$  heights of adjacent tones.

The notation used below is adapted directly from the original work of P80. In that work, the notation “Prominence ( $T$ )” indicated the prominence value of tone  $T$ , while we use  $p(T)$ . Moreover, the notation  $/T/$  was used to refer to the  $f_0$  value of tone  $T$ ; we use  $f_0(T)$ . In addition,  $T^- +$  or  $+T^-$  was used to indicate unstarred tones of bitonal pitch accents while  $T+T$  referred to bitonal pitch accents in which either tone could be starred or unstarred. For example, on pp. 145 of P80,  $H+L$  refers to either  $H^*+L$  or  $H+L^*$ . In the following, we note explicitly which pitch accents are referred to in the expressions of P80. Phonetic rules proposed in P80 are abbreviated in the text as R; for example, Rule 8 is referred to as R8.

#### A.1.1. $L+H^*$ and $L^*+H$ accents

R2, R3, R4, R7, and R8 (pp. 145–146) are relevant to computing  $f_0$  values for the LH tone sequences in  $L+H^*$  and  $L^*+H$  accents. No rule allows for the  $f_0$  values of low and high tones to be computed directly in these contexts; thus, it is necessary to infer the  $f_0$  relationship for LH tone sequences through rule substitution. We begin by observing that R4, shown in (A.1), describes scaling for L in  $L+H^*$  and  $L^*+H$  in the context of a preceding  $H+L^*$ ,  $H^*+L$ ,  $H^*$ , or  $H^*+H$  accent. (Note that  $H^*+H$  was rescinded from the pitch accent inventory for English by Beckman and Pierrehumbert (1986). Moreover, there is apparently a theoretical gap in that no  $f_0$  value can be calculated for L + in  $L+H^*$  when this accent fails to be preceded by a pitch accent with a H tone.)

$$f_0(L) = n f_0(H_i) \frac{p(H_i)}{p(L)} \quad (\text{A.1})$$

Rearranging (A.1) gives the expression:

$$\frac{f_0(L)}{f_0(H_i)} = n \frac{p(H_i)}{p(L)} \quad (\text{A.2})$$

What are the possible values for  $n$ ? Restrictions on values of  $n$  are mentioned in the context of another variable,  $k$ , which is cited in R3, R4, and R8. Note that neither  $n$  nor  $k$  is defined in phonetic terms; these apparently serve merely as mathematical variables modifying prominence values. Of particular relevance, however, for the issue of constraints on these parameters is the statement in R4 that  $0 < n < k$ , as well as the statement in R3 that  $0 < k < 1$ . By transitivity, we can infer that  $0 < n < 1$ , a fact which becomes relevant later in this analysis.

The expression in (A.2) relates the  $f_0$  value of L to the  $f_0$  and prominence values of a *preceding* tone,  $H_i$ , in sequences such as  $H_i+L^*$  or  $L+H_{i+1}^*$ . Because there is no rule for computing the  $f_0$  values of L and H in bitonal  $L+H^*$  or  $L^*+H$  pitch accents, we must instead derive the rule. We observe that R2 relates the scaling of  $H_i$  and  $H_{i+1}$  in sequences such as  $H_i+L^*$  or  $L+H_{i+1}^*$ . R2 is given as

$$f_0(H_{i+1}) = f_0(H_i) \frac{p(H_{i+1})}{p(H_i)}. \quad (\text{A.3})$$

An expression relating the  $f_0$  of L to that of  $H_{i+1}$  is obtained by rearranging (A.3) for  $f_0(H_i)$  and substituting into (A.2), as

$$\frac{f_0(L)}{f_0(H_{i+1})} = n \frac{[p(H_i)]^2}{p(L)p(H_{i+1})}. \quad (\text{A.4})$$

In (A.4) we have obtained an expression relating the  $f_0$  values of adjacent tones L and  $H_{i+1}$ . To understand the syntagmatic relative height restrictions on these two tones, observe that when the left hand side of (A.4)

has a value greater than 1, then L has a higher  $f_0$  than  $H_{i+1}$  in  $L + H_{i+1}^*$  or  $L^* + H_{i+1}$ . Solving for the inequality, it is clear that the  $f_0$  of L is higher than that of  $H_{i+1}$  when  $[p(H_i)]^2/[p(L)p(H_{i+1})] > 1/n$ . To prevent this degenerate situation from occurring, it is necessary to specify that  $[p(H_i)]^2/[p(L)p(H_{i+1})] \leq 1/n$ . However, no such constraint is specified, indicating the result that the relative heights of L and H tones in  $L + H^*$  and  $L^* + H$  are unconstrained, so that the degenerate outputs in Fig. 12(c) and (e) may freely occur. Thus the P80 phonetic model appears to fail to account for the observed relationship in the present experiment between syntagmatic relative tone height and  $f_0$  extremum types and alignment patterns.

A special case which modifies slightly the expression above stating necessary restrictions on the relative heights of L and H tones in  $L + H^*$  and  $L^* + H$  relates to the “downstep rule”, R8. This rule applies when  $L + H_{i+1}^*$  or  $L^* + H_{i+1}$  is preceded by a single-toned  $H^*$  accent.<sup>6</sup> R8 specifies that the expression  $kf_0(H_i)$  should be substituted elsewhere for  $f_0(H_i)$  to account for the lowering of  $f_0$  values of successive H tones. The result of this substitution is to propagate  $k$  to the statement of restrictions on  $f_0$  values, such that L will be higher than adjacent  $H_{i+1}$  precisely when  $[p(H_i)]^2/[p(L)p(H_{i+1})]$  is greater than  $k/n$ . Again, no syntagmatic or other restrictions are given to prevent this degenerate situation from happening, indicating that even when downstep applies, the relative heights of adjacent L and H tones in  $L + H^*$  and  $L^* + H$  are not restricted so as to prevent the degenerate outputs in Fig. 12(c) and (e). Thus the P80 phonetic model appears to fail to account in downstepped contexts as well for the relationship between syntagmatic relative tone height and  $f_0$  extremum types and alignment patterns.

To explore further how L and H may be syntagmatically restricted in other contexts, we can also consider an additional special case involving R7 (p. 145), shown in (A.5). R7 describes the scaling of a “L pitch accent or phrase accent following a  $L^*$  accent,” indicating that this rule is responsible for  $f_0$  contour generation in sequences like  $L^* L^* + H_{i+1}$  and  $L^* + H_{i+1} L^*$ .

$$f_0(L^*) = f_0(L_i^*) \frac{p(L_i^*)}{p(L^*)}. \quad (\text{A.5})$$

What tone scaling restrictions and specifications for tonal sequences of this sort? While R7 specifies how the  $f_0$  value of the *low* tone in  $L^* + H$  is to be calculated, there appears to be no mechanism for computing the  $f_0$  value of *high* tone in the context of a preceding  $L^*$  or  $L^* + H$  accent. It is stated elsewhere (p. 147) that the heights of tones can only be computed with respect to tones as far back as the preceding pitch accent. R2 does not apply, since it describes only how  $+H$  is computed when the preceding context is  $H + L^*$ ,  $H^*$ ,  $H^* + L$  or  $H^* + H$ . As a result of this theoretical gap, the  $f_0$  contour is undefined for  $L^* + H$  in the context of preceding  $L^*$  or  $L^* + H$  accents.<sup>7</sup> The analyses in A.1.1 thus demonstrate that the P80 model fails to account for the relationship between syntagmatic relative tone heights and  $f_0$  extremum characteristics by either (1) generating too many  $f_0$  contours, given some phonological input, or (2) failing to generate any  $f_0$  contour, given some phonological input.

#### A.1.2. $H^* + L$ and $H + L^*$ accents

What syntagmatic restrictions on tone scaling are in place for HL bitonal accents in the P80 model? Consider that R3 and R10 are relevant to computing the  $f_0$  levels of H and L tones in bitonal  $H^* + L$  and  $H + L^*$  accents. R3 (p. 145) is initially presented as giving tonal values for both  $H^* + L$  and  $H + L^*$ . However, it is claimed elsewhere (pp. 159–160) that R3 in fact only applies to  $H + L^*$ , not to  $H^* + L$ . For the  $H^* + L$  pitch accent, L is claimed to be a “floating tone” which is skipped over by phonetic interpolation rules and thus is never phonetically realized with an  $f_0$  value. The fact that the  $f_0$  for  $+L$  is undefined in  $H^* + L$  suggests that the relative  $f_0$  heights of these two tones are also undefined; as a result, these tones’ relative heights cannot be said to be constrained or to correspond to a predictable pattern of  $f_0$  extrema. The revised rule which

<sup>6</sup>Note that Beckman and Pierrehumbert (1986) later rescinded the claim of P80 that HLH sequences triggers downstep, proposing instead that any bitonal pitch accent triggers downstep. However, this proposal has not been supported by corpus data (e.g., Dainora, 2001).

<sup>7</sup>It is not entirely clear whether  $L^* + H$  was meant to be included in the definition of a “L pitch accent” given on p. 145; if not, then the  $f_0$  values of both tones in  $L^* + H$  are undefined in the context of a preceding low pitch accent.

redefines R3 as applying only to  $H+L^*$  is given in R10 (p. 159), shown as

$$f_0(L^*) = kf_0(H+). \quad (\text{A.6})$$

We can consider whether H and L are constrained in  $H+L^*$  on the basis of the expression in (A.6). Although the fact that  $0 < k < 1$  ensures that  $H+$  must be higher than  $L^*$ , no rules specify how the  $f_0$  of  $H+$  can be computed relative to preceding tones. Thus,  $H+L^*$  is phonetically undefined in all positions except phrase-initial position (since the  $f_0$  of the initial tone is always defined in P80). In other words, the P80 model does not generate  $f_0$  contours (or  $f_0$  extrema) from the phonology for the vast majority of HL bitonal contexts.

### A.1.3. $H^*L-$ and $L^*H-$ sequences

What syntagmatic restrictions on adjacent tone scaling are in place for single-toned accent plus phrase accent combinations in the P80 model ( $H^*L-$  or  $L^*H-$ )? R2 and R5 are relevant to computing  $f_0$  values for  $H^*L-$  and  $L^*H-$ . First, consider the relative heights of high and low tones in  $H^*L-$ . R5 (p. 145) applies to this case, shown as

$$f_0(L-) = pf_0(H). \quad (\text{A.7})$$

Given that  $0 < p < k$ , with  $0 < k < 1$ , then by transitivity  $0 < p < 1$ . The  $f_0$  value for  $H^*$  is computable from R2. As a result, the  $f_0$  of  $L-$  is necessarily lower than that of  $H^*$  in  $H^*L-$ .

Next we consider syntagmatic relative height restrictions for  $L^*H-$ . We note another theoretical gap: R2 and R9 state how  $H-$  is to be evaluated in the context of a leftward high tone, but no rules are given for how  $H-$  is to be evaluated in the context of a leftward low tone. Thus, the  $f_0$  contour is undefined for  $L^*H-$ , such that no  $f_0$  contour or associated extrema are generated from the phonetic model for this context.

### A.1.4. Summary

This analysis shows that the syntagmatic relative heights of adjacent H and L tones are constrained for just two out of the six tonal sequences examined in the P80 model:  $H^*L-$  and  $H+L^*$  (when  $H+$  is defined). In contrast, the syntagmatic relative heights of L and H are unconstrained in bitonal  $L+H^*$  and  $L^*+H$ , so that a number of degenerate  $f_0$  shapes and extremum types are possible outputs. Moreover, there are several theoretical gaps for which  $f_0$  values for tones cannot be computed, including  $H-$  in  $L^*H-$ , as well as  $+H$  in  $L^*L^*+H$  or  $L^*+HL^*+H$ . Finally, in  $H^*+L$  the  $+L$  tone is assumed never to be phonetically realized, so that the question of whether  $+L$  can be higher than  $H^*$  does not apply. In sum, there are insufficient syntagmatic restrictions in a number of the rules for scaling the relative heights of adjacent tones in the P80 model, leading to problems with providing a full AM phonological account for the present phonetic data under this model.

## A.2. The Pierrehumbert and Beckman (1988) phonetic model

To what extent are syntagmatic restrictions on relative tone height in the Pierrehumbert and Beckman (1988), or PB88, phonetic model sufficient to support the AM phonological account of the observed relation between relative tone height and  $f_0$  extrema shown in this experiment? The PB88 model is similar to that of P80 in that both H and L tones are assigned values of a prominence parameter which ultimately determines the  $f_0$  values of tones. However, PB88 eliminates the phonetic rules proposed in P80. In the following we describe the model and examine the issue of syntagmatic restrictions on relative tone heights. We note that PB88 used the notation ‘T(H)’ to represent the prominence value of tone H; the notation  $p(H)$  will be used here. Moreover, in PB88 an uppercase letter, e.g. ‘H’, refers to the  $f_0$  value of a H tone; the notation  $f_0(H)$  will be used here.

In PB88 each tone is assigned a prominence value that determines the  $f_0$  value of a tone relative to two phrase-level reference  $f_0$  values, one of which is associated with a high tone line,  $h$ , the other of which is associated with a reference line,  $r$ .  $h$  and  $r$  are given in units of Hz, with  $h > r$ . For a H tone, the  $f_0$  of  $h$  corresponds to a prominence value  $p(H) = 1$  and the  $f_0$  of  $r$  corresponds to a value  $p(H) = 0$ . For an L tone, the reverse is true. (See Fig. 13.) The  $f_0$  values of H and L tones are given by the expressions in

(A.8) and (A.9) below (p. 182):

$$f_0(\text{H}) = [p(\text{H})][h - r] + r, \quad (\text{A.8})$$

$$f_0(\text{L}) = [1 - p(\text{L})][h - r] + r. \quad (\text{A.9})$$

To evaluate the nature of syntagmatic relative height restrictions on adjacent H and L tones in the PB88 model, we first observe that prominence values are assumed not to be influenced by the  $f_0$  and prominence values of neighboring tones. Instead, prominence values are assumed to be determined paradigmatically on the basis of discourse considerations. Since the context-sensitive rules for adjacent tone scaling presented in P80 are eliminated in this model, no sequential syntagmatic dependencies are assumed for prominence or  $f_0$  values of successive H and L tones. In other words, syntagmatic prominence values for adjacent tones are unconstrained. Moreover, values of  $p$  range from 0 to *greater than* 1.0 (p. 186); in other words, the upper limit of  $p$  is unrestricted. Thus, for adjacent H and L, a degenerate  $f_0$  contour will arise when L is higher than H, which occurs precisely when  $p(\text{H}) < 1 - p(\text{L})$ . No restrictions are given to prevent this situation from occurring. In fact, PB88 specifically states (p. 189) that L may rise above an adjacent H in the next accentual phrase, indicating that PB88 intended to permit L to rise higher than H, at least under some circumstances. While PB88 further express doubt that L can rise above H within an accentual phrase, no restrictions are provided limiting the relative heights of L and H tones in any phrasal position. As a result, *the model permits all of the degenerate contours in 12(c)–(f)* and is unable to support an AM phonological account for the present empirical results. An awareness of the possibility that the relative heights of L and H are potentially unconstrained in the phonetic model is indicated by the following quotation (pp. 191–193):

The hierarchical treatment of the  $h$  [the high tone line] seems natural if it is taken to formalize the syntagmatic character of the H/L distinction in Japanese. As long as the minimal boundary prominence is sufficient to place a L% below the phrasal H of its own phrase, every L is lower than the H tones most closely related to it.

Close inspection of PB88 reveals that the claim that “every L is lower than the H tones most closely related to it” is not supported by the formal details of the proposal. Specifically, the claim that restrictions on adjacent L and H tones arise from adjustments of  $h$  under catathesis (i.e., lowering) is inaccurate, since prominence values are not restricted to be lower than the  $h$  tone line. In summary, the relative heights of L and H tones in PB88 are unconstrained for all adjacent tonal sequences, including but not limited to L + H\*, L\* + H, H + L\*, H\* + L, H\*L–, and L\*H–. This means that the AM model cannot, under the assumptions of either the P80 or PB88 models, fully account for the observed relationship between syntagmatic relative tone heights and patterns of  $f_0$  extremum types and alignments.

## References

- Arvaniti, A., Ladd, D. R., & Mennen, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, 26, 3–25.
- Atterer, M., & Ladd, D. R. (2004). On the phonetics and phonology of “segmental anchoring” of  $f_0$ : evidence from German. *Journal of Phonetics*, 32(2), 177–197.
- Bartels, C., & Kingston, J. (1994). Salient pitch cues in the perception of contrastive focus. In P. Bosch, & R. van der Sandt (Eds.), *Focus and natural language processing: Intonation and Syntax*, Vol. 1 (pp. 1–10). Cambridge: Cambridge University Press.
- Beckman, M., & Ayers Elam, G. (1997). *Guidelines for ToBI labeling, version 3.0*. The Ohio State University. <[www.ling.ohio-state.edu/~tobi/ame\\_tobi/annotation\\_conventions.html](http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html)>.
- Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255–309.
- Boersma, P., & Weenink, D. (2002). Praat, a system for doing phonetics by computer (Version 4.0.26): Software and manual available online at <<http://www.praat.org>>.
- Bolinger, D. (1958). A theory of pitch accent in English. *Word*, 14, 109–149.
- Bolinger, D. (1961). Contrastive accent and contrastive stress. *Language*, 37, 83–96.
- Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: Gleerups.
- Caspers, J., & van Heuven, V. J. (1993). Effects of time pressure on the phonetic realization of Dutch accent-lending pitch rise and fall. *Phonetica*, 50, 161–171.

- Chen, A. (2003). Reaction time as an indicator of discrete intonational contrasts in English. In *Proceedings of Eurospeech* (pp. 97–100). Geneva.
- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row (Reprinted 1991, Boston: MIT Press).
- Dainora, A. (2001). *An empirically based probabilistic model of intonation in American English*. Ph.D. dissertation, University of Chicago.
- Dilley, L. C. (2005). *The phonetics and phonology of tonal systems*. Ph.D. dissertation, MIT, Cambridge, MA.
- Dilley, L. C. (2006). Looking beneath the surface: Why AM theory needs to be revised. Poster presented at the 10th Laboratory Phonology Conference, Paris, France.
- Dilley, L. C. (to appear). On the dual relativity of tone. *Proceedings of the 41st annual regional meeting of the Chicago Linguistics Society*.
- Dilley, L. C., Ladd, D. R., & Schepman, A. (2005). Alignment of L and H in bitonal pitch accents: testing two hypotheses. *Journal of Phonetics*, 33(1), 115–119.
- D'Imperio, M. (2000). *The role of perception in defining tonal targets and their alignment*. Ph.D. dissertation, The Ohio State University.
- Gili Fivela, B. (in press). The coding of target alignment and scaling in pitch accent transcription. *Italian Journal of Linguistics*.
- Goldsmith, J. (1976). *Autosegmental phonology*. Ph.D. dissertation, MIT, Cambridge, MA.
- Grice, M., Ladd, D. R., & Arvaniti, A. (2000). On the place of phrase accents in intonational phonology. *Phonology*, 17, 143–185.
- Grice, M. & Savino, M. (1995). Low tone versus 'sag' in Bari Italian intonation: A perceptual experiment. In *Proceedings of the international congress of phonetic sciences* (pp. 658–661). Stockholm.
- Gussenhoven, C. (1999). Discreteness and gradience in intonational contrasts. *Language and Speech*, 42, 283–305.
- Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge University Press.
- Gussenhoven, C., & Rietveld, A. C. M. (1988). Fundamental frequency declination in Dutch: Testing three hypotheses. *Journal of Phonetics*, 16, 355–369.
- Gussenhoven, C., & Rietveld, T. (2000). The behavior of H\* and L\* under variations in pitch range in Dutch rising contours. *Language and Speech*, 43(2), 183–203.
- Gussenhoven, C., Repp, B., Rietveld, A., Rump, W., & Terken, J. (1997). The perceptual prominence of fundamental frequency peaks. *Journal of the Acoustical Society of America*, 102, 3009–3022.
- Halliday, M. A. K. (1967). *Intonation and grammar in British English*. The Hague: Mouton.
- Hirschberg, J., & Ward, G. (1992). The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics*, 20, 241–251.
- House, D. (1990). *Tonal perception in speech*. Lund: Lund University Press.
- Knight, R.-A. (2003). *Peaks and plateaux: The production and perception of high intonational targets in English*. Ph.D. dissertation, University of Cambridge, Cambridge.
- Kohler, K. J. (1987). Categorical pitch perception. In: U. Viks (Ed.), *Proceedings of the 11th International Congress Of Phonetic Sciences* (pp. 331–333). Vol. 5. Tallinn.
- Ladd, D. R. (1990). Metrical representation of pitch register. In J. Kingston, & M. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 35–57). Cambridge: Cambridge University Press.
- Ladd, D. R. (1993). In defense of a metrical theory of intonational downstep. In H. van der Hulst, & K. Snider (Eds.), *The phonology of tone: The representation of tonal register* (pp. 109–132). Berlin, New York: Mouton de Gruyter.
- Ladd, D. R. (1994). Constraints on the gradient variability of pitch range, or, pitch level 4 lives!. In P. A. Keating (Ed.), *Papers in laboratory phonology III: Phonological structure and phonetic form* (pp. 43–63). Cambridge, MA: Cambridge University Press.
- Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.
- Ladd, D. R. (2000). Tones and turning points: Bruce, Pierrehumbert, and the elements of intonational phonology. In M. Horne (Ed.), *Prosody: Theory and experiment—Studies presented to Gösta Bruce* (pp. 37–50). Dordrecht: Kluwer.
- Ladd, D. R., Faulkner, D., Faulkner, H., & Schepman, A. (1999). Constant “segmental anchoring” of  $f_0$  movements under changes in speech rate. *Journal of the Acoustical Society of America*, 106(3), 1543–1554.
- Ladd, D. R., Mennen, I., & Schepman, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America*, 107(5), 2685–2696.
- Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics*, 25, 313–342.
- Ladd, D. R., & Schepman, A. (2003). “Sagging transitions” between high accent peaks in English: Experimental evidence. *Journal of Phonetics*, 31, 81–112.
- Ladd, D. R., Verhoeven, J., & Jacobs, K. (1994). Influence of adjacent pitch accents on each other's perceived prominence. *Journal of Phonetics*, 22, 65–85.
- Lieberman, M. (1975). *The intonational system of English*. Ph.D. dissertation, MIT, Cambridge, MA.
- Lieberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff, & R. Oerhle (Eds.), *Language sound structure* (pp. 157–233). Cambridge, MA: MIT Press.
- Lickley, R. J., Schepman, A., & Ladd, D. R. (2005). Alignment of phrase accent low in Dutch falling rising questions: Theoretical and methodological implications. *Language and Speech*, 48, 157–183.
- Moore, B. C. J. (2003). *An introduction to the psychology of hearing*. New York: Academic Press.
- Moulines, E., & Charpentier, F. (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9, 453–467.
- Niebuhr, O. (2003). Perceptual study of timing variables in  $f_0$  peaks. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1225–1228). Barcelona.
- Nolan, F. (2003). Intonational equivalence: An experimental evaluation of pitch scales. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 771–774). Barcelona.

- O'Connor, R. J., & Arnold, G. F. (1973). *Intonation of Colloquial English*. Bristol, UK: Longman Group, Ltd.
- Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Ph.D. dissertation, MIT, Cambridge, MA.
- Pierrehumbert, J., & Beckman, M. (1988). *Japanese tone structure*. Cambridge, MA: MIT Press.
- Pierrehumbert, J., & Steele, S. A. (1989). Categories of tonal alignment in English. *Phonetica*, 46, 181–196.
- Pike, K. L. (1945). *The Intonation of American English*. Ann Arbor: University of Michigan Press.
- Prieto, P., D'Imperio, M., & B. Gili-Fivela (2005). Pitch accent alignment in Romance: Primary and secondary associations with metrical structure. *Language and Speech*, 48, 359–396.
- Redi, L. (2003). Categorical effects in the production of pitch contours in English. In *Proceedings of the 15th International Congress of the Phonetic Sciences* (pp. 2921–2924), Barcelona.
- Rolland, G., & Lævenbruck, H. (2002). Characteristics of the accentual phrase in French: An acoustic, articulatory, and perceptual study. In *Proceedings of Speech Prosody 2002* (pp. 611–614). Aix-en-Provence, France.
- Shattuck-Hufnagel, S., Dilley, L., Veilleux, N., Brugos, A., & Speer, R. (2004).  $f_0$  peaks and valleys aligned with non-prominent syllables can influence perceived prominence in adjacent syllables. In *Proceedings of Speech Prosody 2004*. Nara, Japan.
- Silverman, K. E. A., & Pierrehumbert, J. (1990). The timing of prenuclear high accents in English. In J. Kingston, & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 71–106). Cambridge, UK: Cambridge University Press.
- Snider, K. (1999). *The Geometry and Features of Tone*. Dallas: Summer Institute of Linguistics and the University of Texas at Arlington Publications in Linguistics 133.
- Snider, K., & van der Hulst, H. (1992). Issues in the representation of tonal register. In H. van der Hulst, & K. Snider (Eds.), *The phonology of tone: The representation of tonal register* (pp. 1–27). Berlin: Mouton de Gruyter.
- 't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation*. Cambridge: Cambridge University Press.
- Vanrell Bosch, M. M. (2006). A scaling contrast in Majorcan Catalan interrogatives. In *Proceedings of Speech Prosody 2006*, Dresden.
- Ward, G., & Hirschberg, J. (1985). Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 61, 747–776.
- Welby, P. (2003). *The slaying of Lady Mondegreen, being a study of French tonal association and alignment and their role in speech segmentation*. Ph.D. dissertation, The Ohio State University.
- Welby, P. (2006). French intonational structure: Evidence from tonal alignment. *Journal of Phonetics*, 34, 343–371.
- Williams, E. S. (1971/6). Underlying tone in Margi and Igbo. *Linguistic Inquiry*, 7, 463–484 [Ms. written in 1971].
- Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25, 61–83.
- Xu, Y. (1998). Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, 55, 179–203.
- Xu, Y. (1999). Effects of tone and focus on the formation and alignment of  $f_0$  contours. *Journal of Phonetics*, 27, 55–105.
- Xu, Y. (2001). Fundamental frequency peak delay in Mandarin. *Phonetica*, 58, 26–52.
- Xu, Y. (2002). Articulatory constraints and tonal alignment. In *Proceedings of the 1st International Conference on Speech Prosody* (pp. 91–100). Aix-en-Provence, France.
- Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46, 220–251.
- Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America*, 111, 1399–1413.
- Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33, 319–337.
- Xu, C. X., & Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association*, 33, 165–181.
- Xu, Y., & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, 33, 159–197.
- Xu, Y., Xu, C.X., & Sun, X. (2004). On the temporal domain of focus. In *Proceedings of the international conference on speech prosody 2004* (pp. 81–84). Nara, Japan.