The Phonetic Specification of Contour Tones: Evidence from the Mandarin Rising Tone

Edward Flemming[a] and Hyesun Cho[b]

[a]*Massachusetts Institute of Technology,* [b]*Dankook University,*

ABSTRACT

This paper investigates the phonetic specification of contour tones through a case study of the Mandarin rising tone. The patterns of variation in the realization of the rising tone as a function of speech rate indicate that its specifications include targets pertaining to both the pitch movement and its end points: the slope of the f0 rise, the magnitude of the rise, and the alignment of the onset and offset of the rise. This analysis implies that the rising tone is over-specified in that any one of the target properties can be derived from the other three (e.g. slope is predictable from the magnitude and timing of the rise). As a result, the targets conflict and cannot all be realized. The conflict between tone targets is resolved by a compromise between them, a pattern that is analyzed quantitatively by formulating the targets as weighted, violable constraints.

1. Introduction

In principle, the pitch movement for a contour tone can be characterized in terms of a variety of properties. For example, a rising f0 movement of the kind illustrated schematically in figure 1 can be described in terms of the timing of the onset (*L*) and offset (*H*) of the rise with respect to landmarks (or 'segmental anchors') in the segmental string (labeled A_L and A_H), the pitch levels of *L* and *H*, the magnitude, *M*, of the rise in f0, and the average slope of the rise (*S*), among others. Many of these properties can be derived from each other. For example, the magnitude of the rise, *M*, is equal to the difference in the pitch levels of *L* and *H*, and the average slope, *S*, is equal to the magnitude of the rise divided by the duration between *L* and *H*. Accordingly only a subset of these properties need to be specified to specify a rising f0 movement, and theories of the phonetic implementation of tones posit that contour tones are specified in terms of targets for various subsets of these properties.

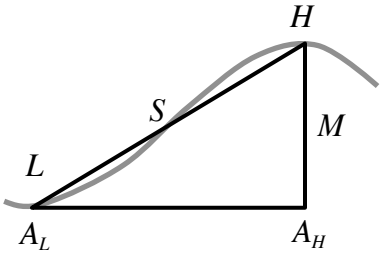

Fig. 1. A schematic illustration of a rising F_0 movement. *L* and *H* are the onset and offset of the f0 movement, A_L and A_H mark positions in the segmental string that serve as segmental anchors for *L* and *H*, *M* is the magnitude of the rise in F_0, and *S* is its average slope.

For example, many models of tonal realization follow the general outlines of Pierrehumbert's (1980) analysis of English intonation. Pierrehumbert adopts the standard phonological analysis according to which contour tones are composed of sequences of level tone specifications (Goldsmith 1976). Each tone unit, whether a simplex tone or part of a contour tone, is hypothesized to be realized by a level target specified for fundamental frequency and alignment with respect to the segmental string. Transitions between these level targets are derived by general interpolation mechanisms so properties of transitions, such as their slopes, are not regulated by phonetic targets (Pierrehumbert 1980:47-52). This claim is explicitly articulated by Ladd and colleagues as part of the Segmental Anchoring Hypothesis, according to which 'the beginning and end of a pitch movement are anchored to specific locations relative to segmental structure, while the slope and duration of the pitch movement vary according to the segmental material with which it is associated' (Ladd 2004)[1]. This approach has been successfully applied to the analysis of the realization of rising pitch accents in languages such as Greek (Arvaniti, Ladd & Mennen 1998).

However the slope of the F_0 trajectory during a tone is an important cue to the distinction between level and rising tones in a language like Mandarin (Gandour 1979, 1984, Massaro, Cohen & Tseng 1985), so we might expect such tones to have a specified target for the slope of the transition, contrary to the Segmental Anchoring Hypothesis. For example, a rising tone could be specified in terms of alignments of *L* and *H*, the slope of the rise and the pitch level of *L*. The magnitude of the rise and the pitch level of *H* would then follow from the slope and duration of the rise. Models of Mandarin tone realization that incorporate targets pertaining to the shapes of pitch movements have been proposed by Kochanski, Shih & Jing (2003) and Xu & Wang (2001). Slope targets have also been proposed in the context of a model of intonation by 't Hart et al (1990: 72-77).

This study investigates the phonetic specification of the Mandarin rising tone (tone 2), testing for targets for three of the properties illustrated in figure 1: (i) the alignments of *L* and *H* to the segmental string, (ii) magnitude, *M*, and (iii) slope, *S*. Due to the difficulty of normalizing pitch range across speakers, pitch levels are not analyzed here, so we will not distinguish between an actual magnitude target as opposed to targets for the pitch levels of *L* and *H* (but see section 5.2 for further discussion). The nature of the tonal targets is investigated by examining the realization of the rising tone under variation in time pressure (cf. Caspers and van Heuven 1993). If the rising tone has targets for a subset of the properties under investigation then those properties should not vary as a function of speech rate, whereas unspecified properties should vary systematically with speech rate.

For example, if the onset and offset of the rise are consistently aligned to segmental anchors such as the middle and end of the syllable, then as speech rate increases, moving the anchors closer together, the duration of the rise should decrease. If speakers also try to maintain a target magnitude of F_0 rise, then a decrease in rise duration will result in increased slope (Fig. 2(a)). On the other hand, if there is a constant slope target but magnitude is unspecified, then an increase in speech rate will result in a rise of smaller magnitude (Fig. 2(b)). If speakers try to keep both slope and magnitude constant, then the duration of the rise must be constant and consistent segmental anchoring will not be possible, so alignment of *L* and *H* with respect to the segmental string should vary as a function of speech rate, with

[1] Pierrehumbert's (1980) analysis of English intonation is not completely consistent with the Segmental Anchoring Hypothesis because the trailing tone of a complex pitch accent, like the H of L*+H, is said to be timed at a fixed interval following the L tone rather than being aligned to a segmental anchor (pp. 77-80).

L occurring earlier in the syllable and/or *H* occurring later in the syllable as syllable duration gets shorter (Fig. 2(c)). Thus the patterns of variation in tone realization as speech rate changes can reveal the nature of the targets of the rising tone.

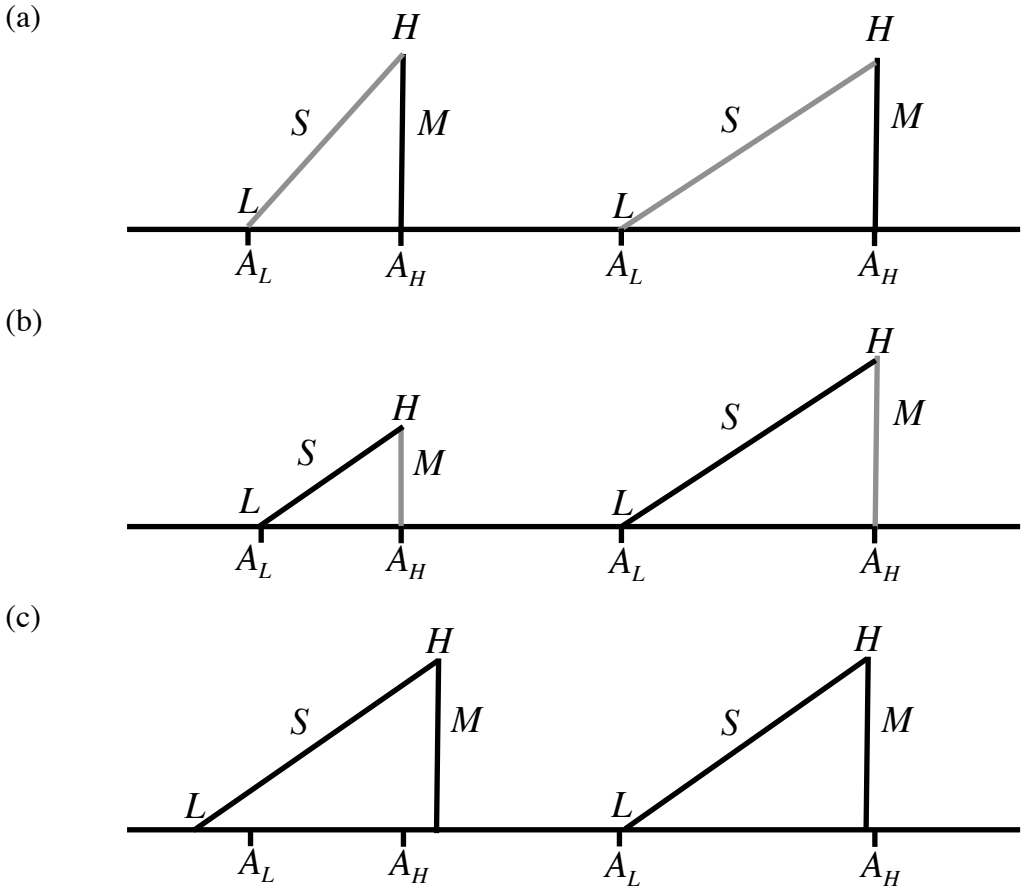

Fig. 2. Schematic illustration of the predicted effects of variation in time pressure on the realization of a rising tone. The left of each panel represents a faster speech rate than the right, so the segmental anchors A_L and A_H are closer together. (a) *L* and *H* are consistently aligned to their respective segmental anchors and rise magnitude, *M*, is constant so slope increases as speech rate increases. (b) *L* and *H* are consistently aligned to their respective segmental anchors and slope, *S*, is constant so magnitude *M* decreases with increasing speech rate. (c) slope *S* and magnitude *M* are constant, so alignment of *L* and *H* with respect to anchors A_L and A_H varies as a function of speech rate.

We will see that in fact all of these properties of the Mandarin rising tone vary as a function of speech rate, so it is not possible to make a division into targeted and contingent properties of the tone. Rather, the patterns of variation can be accounted for by positing violable targets for all of the properties under consideration. These results imply that a model based on point targets plus general interpolation mechanisms cannot provide an adequate account of tonal realization – it is necessary to allow for targets that pertain to the properties of the transition between the end-points of a contour tone, such as the slope of the transition. However, the results also show that it is not possible to substitute

transition targets for any of the targets pertaining to the end-points of the rise: the slope target is required in addition to targets for the end-points.

Specifying targets for *L, H, M* and *S* involves a form of 'over-specification' since the targets generally cannot all be realized. For example, as noted above, the slope, *S*, can be calculated from rise magnitude, *M*, together with the timing of *L* and *H*, so a slope target generally conflicts with targets for tone alignment and *M*. We will see that the patterns of variation in the realization of the rise as a function of speech rate can be analyzed as resulting from a compromise between the conflicting demands of these targets. This analysis has implications that extend beyond the specific case of contour tones because it implies that phonetic realization can operate in terms of violable targets, and that it incorporates mechanisms for the resolution of conflicts between targets. Specifically, we show that the analysis can be formalized in terms of a model according to which phonetic realizations are optimized with respect to weighted, potentially conflicting constraints (Flemming 2001).

2. Experiment

Mandarin contrasts four full tones: high level (tone 1), rising (tone 2), low falling-rising (tone 3) and falling (tone 4), illustrated as spoken on isolated monosyllables in figure 3. This study examines the rising tone (2). The goal was to investigate the realization of this tone as the duration of the syllable bearing the tone varies. The duration of the syllable with which the tones were associated was varied by selecting syllables whose segments varied in inherent duration – e.g. high and low vowels – and by eliciting them at three different speech rates.

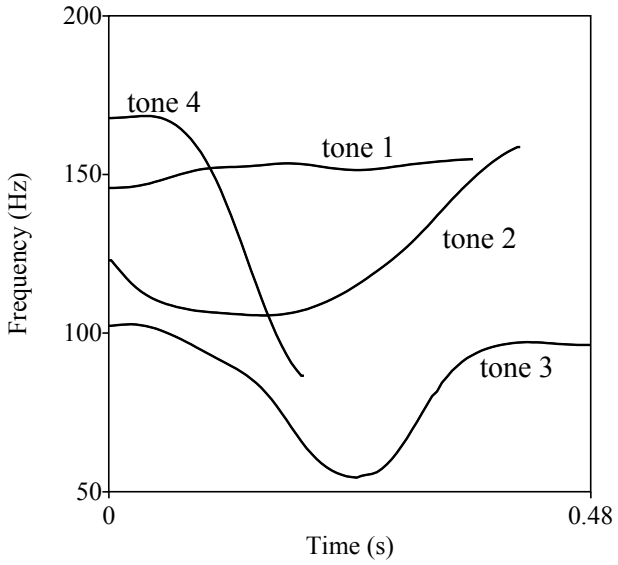

Fig. 3. Pitch tracks of the four Mandarin tones, produced on isolated words (recordings from http://www.phonetics.ucla.edu/vowels/chapter2/chinese/recording2.1.html).

2.1. Speech materials

The main materials consist of 14 target syllables, each bearing a rising tone. These target syllables were initial in disyllabic words where the second syllable also bears a rising tone (1(a)). All words

consist entirely of sonorant sounds. The full materials also included one additional word of the same form but beginning with a voiced stop, which was excluded from analysis due to strong effects of the stop on fundamental frequency. There were five additional words in which a rising tone is followed by a neutral-toned syllable (1(b)), discussed in section 5.2.2, and nine fillers consisting of three and four syllable words with various combinations of tones. All words were produced in the carrier phrase [tɕʰi̯ŋ nǐn pà __ tsâi ʂɔ́ jǐ bjɛ̀n] ("Please say ____ again"), so the target tones in (1(a)) were preceded by a low tone and followed by a rising tone.

(1) List of target words in Pinyin with IPA transcription and English gloss.

(a) Rising tone followed by rising tone

míngnián	[mǐŋnjɛ̌n]	'next year'
míngrén	[mǐŋɻə̌n]	'celebrity'
línglíng	[lǐŋlǐŋ]	'cool'
míngmíng	[mǐŋmǐŋ]	'clearly'
niánlíng	[njɛ̌nlǐŋ]	'age'
nóngmín	[nɔ̌ŋmǐn]	'farmer'
rénmín	[ɻə̌nmǐn]	'people'
ménlíng	[mə̌nlǐŋ]	'doorbell'
léiléi	[lěilěi]	'hang in clusters'
nánnán	[nănnăn]	'murmuring'
lángláng	[lɑ̌ŋlɑ̌ŋ]	'clear and ringing'
mángrén	[mɑ̌ŋɻə̌n]	'blind person'
láilín	[lăilǐn]	'arrive'
láinián	[lăinjɛ̌n]	'next year'

(b) Rising tone followed by neutral tone

míngzi	[mǐŋtsɨ]	'name'
piányi	[pjɛ̌nji]	'cheap'
rénmen	[ɻə̌nmən]	'people'
péngyou	[pʰə̌ŋjoʊ]	'friend'
zánmen	[tsǎnmən]	'we'

2.2. Participants and procedure

The subjects were four native speakers of Beijing Mandarin Chinese, two male (one in his 40's, the other in his 60's) and two female (one in her 20's, the other in her 60's). They were paid for their participation in the study.

The sentences were presented on paper in a randomized order. Subjects were asked to read all of the materials twice at each of three speech rates: first at a self-selected rate, then at a faster speech rate and finally at a slower rate. Recordings were made in a sound-attenuated booth using a Shure SM10A close-talking microphone. The signal was pre-amplified and digitized at 44.1 kHz, 16 bits per sample, using a USBPre microphone interface and recorded direct to disk on an iMac running Amadeus II software (version 3.8.4, Hairersoft).

Two speakers produced one item (njěnlǐŋ 'age') with a falling or neutral tone on the second syllable rather than the expected rising tone, so these utterances were excluded from analyses. Nine additional utterances were discarded due to problems with pitch tracking, leaving a total of 315 rising tone tokens in the main analyses.

2.3. Measurements

The basic measurements taken from each token were the timing of segment boundaries and the timing and level of onset (*L*) and offset (*H*) of the F_0 rise associated with the rising tone and the F_0 levels at these times. All times were measured relative to word onset.

Segment boundaries were labeled by hand in *Praat* (Boersma & Weenink 2009), with reference to spectrograms and waveforms. Boundaries between consonants and vowels were straightforward to identify based on rapid changes in intensity, accompanied by abrupt spectral changes. Boundaries between consonants in words like [mǐŋ.mǐŋ]) were often marked by visible release of the first consonant or abrupt spectral changes, but were hard to identify in some cases. As discussed below, the temporal mid-point of the intervocalic cluster proved to be more relevant to the timing of F_0 events than the release of the first consonant, so we did not have to rely on these measurements.

In many studies of rising pitch accents, the onset and offset of the rise are taken to correspond to the F_0 minimum and maximum respectively. Here we adopt the assumption that the F_0 maximum corresponds to the offset of the rise, *H*, (with a few exceptions discussed below), but the F_0 minimum does not appear to be a significant F_0 event in the realization of the Mandarin rising tone. This is because the F_0 trajectory of the rising tone often shows a relatively level interval followed by a rise (Fig. 4), and the timing of the F_0 minimum varies substantially as a function of minor variations in the realization of the low plateau. For example, in Fig. 4(a) the F_0 minimum precedes the target syllable because the low plateau rises slightly, whereas in Fig. 4(b) it occurs much later because the low plateau falls slightly. The low plateau at the beginning of the rising tone is plausibly analyzed as the result of interpolation between a low target at the offset of the preceding low tone and a low target at the onset of the final rise. This interpretation is supported by the fact that F_0 falls steadily to onset of the final rise when a rising tone is preceded by a high tone (Xu 1997:69, Chen & Gussenhoven 2008:730) – both patterns can be accounted for by smooth interpolation between the last target of the preceding tone and the low target at the onset of the final rise in the rising tone. Accordingly, we take the F_0 event corresponding to *L* to be the onset of the rapid final rise (cf. Xu 1998:196f., Chen & Gussenhoven 2008:730). This time-point was identified algorithmically, as described below.

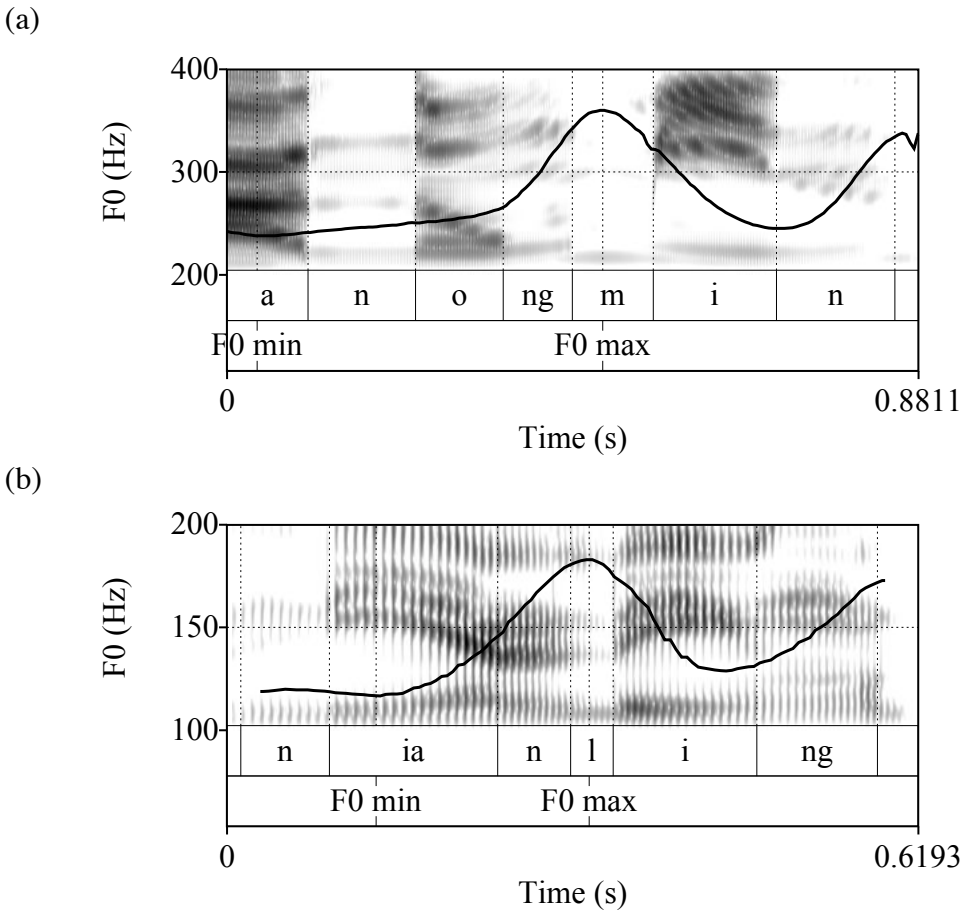

Fig. 4. Pitch tracks and spectrograms of (a) [nǒŋmǐn] 'farmer' and (b) [njěnlǐŋ] 'age'.

Although *H* was identified as the local maximum in F_0 in the great majority of cases, there were six utterances, produced at fast speech rate, where the second rising tone was so reduced that it was produced without any local minimum in F_0. Instead the F_0 rise slowed after the first tone, but continued to rise, peaking late in the second syllable, as in Fig. 5. In these cases, it is clear that the F_0 maximum belongs to the second rising tone, so *H* is marked at the 'shoulder' where F_0 becomes level, or close to level, after the initial rise.

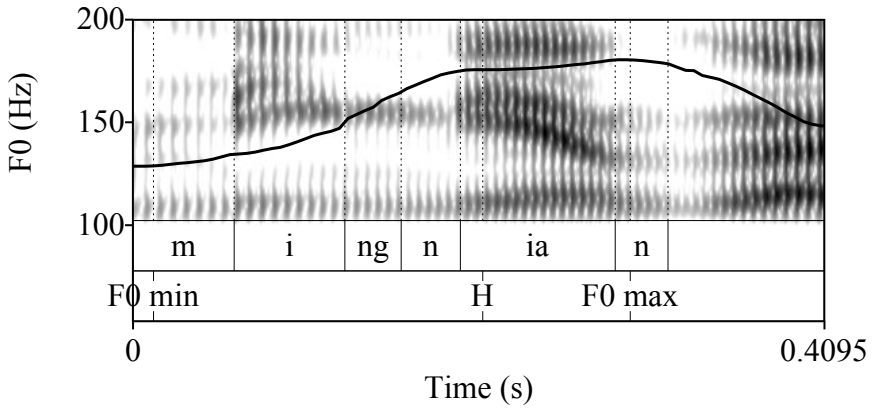

Fig. 5. Pitch track and spectrograms of [mi̍ŋnjɛ̍n] 'next year' produced at fast speech rate. The second rising tone is realized as a high plateau followed by a slight rise in F_0.

The identification of the onset of the rise involves identifying inflection points or 'elbows' in the F_0 trajectories, which is a notoriously difficult problem (e.g. del Giudice et al 2007). We labeled the onset of the rise, *L*, using two procedures, one simple but relatively crude, and the other more complex. Analyses were conducted based on both measures to assess the robustness of the results with respect to details of the definition of the *L* time point. We will see that they yield qualitatively identical patterns of results, indicating that those results are not artifacts of a particular algorithm for identifying *L*.

The first approach adapts a procedure described in D'Imperio (2000) and Welby (2006), and endorsed by del Giudice et al (2007): the trajectory from F_0 minimum to F_0 maximum was approximated by three straight lines, fitted to minimize squared error. The best fitting approximation was found by testing every possible division of a sampled F_0 trajectory into three segments, sampling at 5 ms intervals, with a minimum segment length of four samples. 'Elbows' were then taken to correspond to the time points at which adjacent line segments intersected, and the onset of the rise was taken to be the elbow that begins the steepest line segment. This procedure is illustrated in Fig. 6, where *L* is identified as the start of the second line segment. In one case the second and third lines intersected before the first and second. This utterance was discarded from analyses of *L* defined in terms of piecewise linear approximation.

A three-line approximation was used rather than the two lines proposed by D'Imperio (2000) because the elbow location determined by a two-line approximation is very sensitive to the global shape of the trajectory. For example, if the high peak is relatively broad, the trajectory can be best approximated by a steeply rising segment followed by a shallow segment at the peak, putting the onset of the rise at the beginning of the trajectory, while the same shape of rise with a narrower peak may be divided into a low plateau followed by a steep rise, resulting in different locations for the onset of the rise depending on the shape of the offset of the rise. The three-line approximation allows for the possibility of shallow segments at the beginning and end of a trajectory, as in Fig. 6.

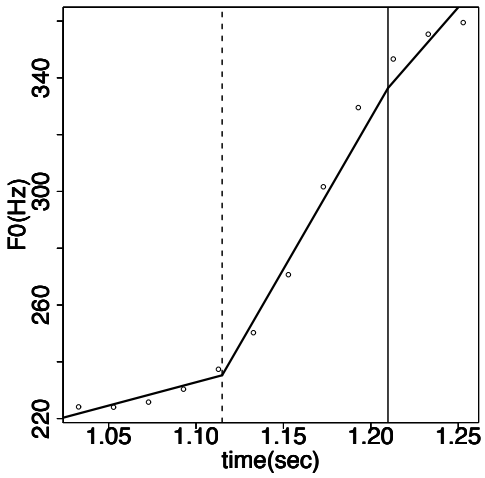

Fig. 6. Example of a three-line approximation to the trajectory between F_0 minimum and maximum. F_0 measurements are plotted with open circles and the intersections of the line segments are marked by vertical lines. The onset of the final rise, *L*, is identified as the beginning of the steepest line segment, marked by the dashed vertical line.

A problem with this first method for identifying the onset of the rise is that it is designed to separate a low plateau from a following rise, but rising tones are sometimes realized without any identifiable low plateau, particularly at fast speech rates. In these cases the onset of the final rise should be close to the F_0 minimum, but the algorithm divides the rise into three parts anyway which often results in the onset being placed rather late in the rise. For this reason, we developed a second algorithm to identify the onset of the rise based on the velocity and acceleration of the F_0 trajectory. Velocity and acceleration were calculated by fitting a fourth order natural smoothing spline to the F_0 trajectory from 10 ms before the F_0 minimum to 20 ms after the F_0 maximum, using the function *smooth.Pspline* from the R package *pspline* (Ramsay & Ripley 2013), with smoothing parameter set to 10^{-11}, then calculating derivatives of the smoothed curve.

The starting point for this approach is the observation that the onset of the final rise could be identified as the point at which F_0 velocity rises above zero if the low plateau were always perfectly flat or falling. While examples like Fig. 4(a) show that the plateau can in fact rise, its velocity is relatively low and steady compared to the final rise, so identifying the onset of the rise as the point where velocity exceeds a threshold of 20% of peak velocity generally avoids misidentifying a rising plateau as part of the final rise while remaining close to visually identifiable elbows in the F_0 trajectory. More precisely, the onset of the rise is taken to be the point at which velocity exceeds 20% of maximum and remains above that level until the velocity peak, because brief rises above the 20% threshold are generally due to segmental perturbations.

This velocity-based criterion fails in a small, but non-negligible, number of examples where a rising plateau exceeds 20% of peak velocity but there is still a clear elbow marking the onset of the final rise, so the performance of this algorithm can be improved by taking F_0 acceleration into account as well. A plateau has relatively constant velocity even if it is rising, so the transition to the final rise is marked by a rise in acceleration as well as velocity, with acceleration reaching a local maximum before the point

of peak velocity, as in Fig. 7[2]. Accordingly, we identify the onset of the rise as the first point after velocity exceeds 20% of its maximum where acceleration also exceeds 25% of its local maximum. This point is marked by the vertical dotted line in Fig. 7. However two qualifications have to be added: (i) Segmental perturbations of F_0 can also give rise to local maxima in acceleration, however these local maxima are usually smaller than maxima associated with an elbow in the trajectory, so an acceleration maximum is only taken to indicate the presence of an elbow if it exceeds a preceding acceleration minimum by at least 1500 Hz/s/s. (ii) If no such acceleration maximum is observed, that generally indicates that there is no plateau present, or that the plateau ends close to the F_0 minimum, so the rise onset is identified based on the velocity criterion alone (i.e. the last point where velocity exceeds 20% of its maximum). These thresholds were selected to yield a good match between visually obvious elbows in F_0 trajectories and the time points identified by the algorithm. Relatively high thresholds were required to avoid misidentifying segmental perturbations of F_0 as the onset of the rise.

The results of the two algorithms are highly correlated ($r = 0.98$), although the linear approximation algorithm almost always placed *L* later than the algorithm based on velocity and acceleration of F_0, with a mean difference of 39 ms. For reasons of space, only analyses based on *L* derived from the velocity/acceleration algorithm are reported below, but exactly the same patterns were observed in statistical analyses of the linear-approximation *L* measure, indicating that these results are robust to the details of the algorithm used to identify the onset of the rise.

Having located *L* and *H*, the magnitude *M* is then calculated from F_0 at *H* minus F_0 at *L*, and the average slope is *M* divided by the duration from *L* to *H*. Peak velocity was also measured from the smoothed velocity curve. We focus on average slope since its mathematical relationship to *M*, *L* and *H* slightly simplifies the model proposed in section 4.2, but it is highly correlated with peak velocity ($r = 0.98$), so the two are largely interchangeable in the analyses that follow.

[2] In a study of the Mandarin rising tone, Xu (1998) identifies the onset of the rise as the point of maximum acceleration in F_0, but in a smooth rise, the acceleration maximum occurs well after the onset of the rise, and this is often the case in our data (cf. Li 2003:45).

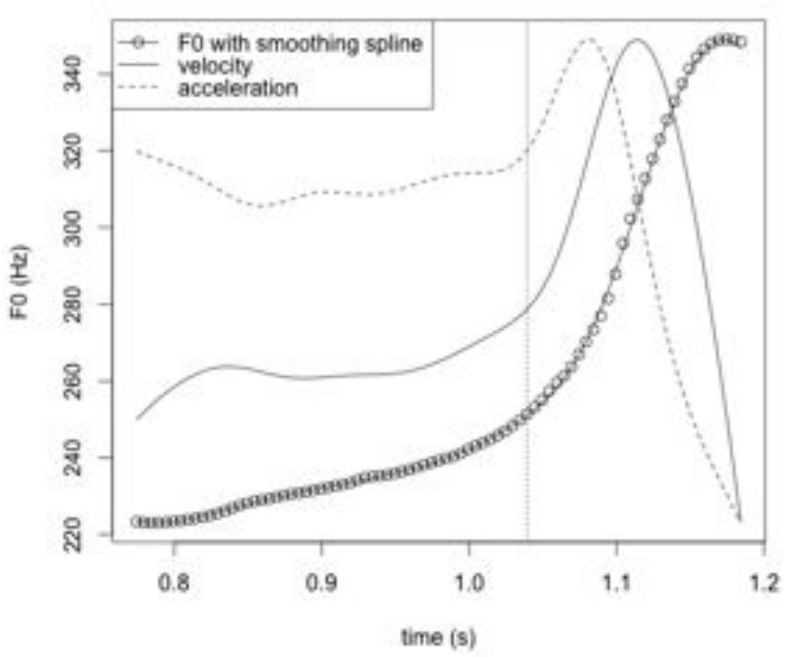

Fig. 7. Trajectory between F_0 minimum and maximum (plotted with open circles), smoothed with 4^{th} order natural smoothing splines, plotted with velocity (solid line) and acceleration curves (dashed line) calculated from the smoothed F_0. The vertical dotted line marks the onset of the final rise, as identified by the algorithm described in the text.

3. Results

The immediate goal of the experiment is to test whether the various properties of the rising tone do or do not vary systematically as a function of segmental durations. We first investigate the timing of the onset, *L*, and offset, *H* of the rise, then turn to the magnitude and slope of the rise.

3.1. Segmental anchoring

The claim that the onset and offset of the rising tone should be invariantly aligned to fixed segmentally-defined locations, regardless of segmental durations, is part of the Segmental Anchoring Hypothesis (Ladd 2004). To investigate this hypothesis with respect to the Mandarin rising tone, we conducted a search for segmentally defined points that are consistently aligned with *L* and *H* respectively across variations in speech rate and segmental make-up of the syllable. Based on previous work on tonal alignment in general and on the alignment of the Mandarin rising tone in particular, we identified a set of candidate segmental anchors of two types: segment boundaries, such as the onset of the vowel, and proportions of phonological constituents, such as the middle of the syllable. The full set is listed in (2), and the relevant time points are illustrated in Fig. 8.

(2) Candidate segmental anchors for *L* and *H*

Segment boundaries:
- Onset of the syllable (C1)
- Onset of the vowel (V1)
- Offset of the vowel (C2 or Off)
- Offset of the syllable (Off)
- Onset of the following vowel (V2)

Proportions of the following constituents:
- Syllable (C1 to Off)
- Rime (V1 to Off)
- Vowel-to-vowel interval (V1 to V2)

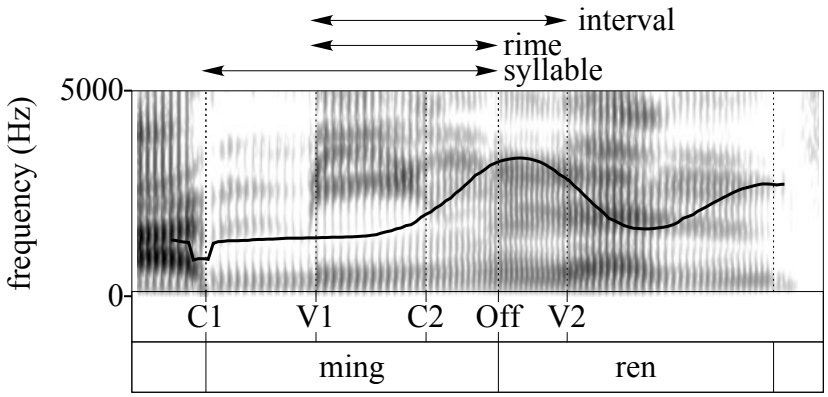

Fig. 8. Pitch track and spectrogram of a bisyllabic target word showing the locations of the segment boundaries and constituents listed in (2).

The syllable is taken to begin at the formation of the constriction for the onset consonant so a CV syllable extends from the beginning of C1 to the formation of the constriction for the onset of the following syllable (labeled 'Off' in Fig. 8). As noted above, the offset of a CVC syllable can be difficult to identify because it is located in the middle of a -CC- cluster (e.g. [mĭn.mĭŋ]), and the precise location of the closure for the second consonant is sometimes unclear from the acoustic signal. In the analyses of tone timing it turned out that locating the offset of CVC syllables at a point halfway through the duration of the -CC- sequence yielded better results than trying to identify the consonant boundary, so that is the criterion adopted in the analyses reported below. The offset of the vowel coincides with the syllable offset in CV syllables, and with the onset of C2 in CVC syllables. The rime extends from the onset of the vowel, V1, to the offset of the syllable, Off. The vowel-to-vowel interval (Farnetani & Kori 1986, Steriade 2012) extends from the onset of the first vowel, V1, to the onset of the vowel of the following syllable, V2, so in the word [lěiliěi] the relevant interval is the sequence [ěil], while in the word [mĭŋmĭŋ] it is the sequence [ĭŋm].

The candidate segmental anchors were evaluated by comparing the goodness of fit of models predicting the timing of *L* and *H* based on each anchor. If a given segmental boundary is the anchor for a tone then that tone should always occur at that boundary, or at a short fixed interval preceding or following it. This implies a model of tone timing of the form shown in (3), where *T* is the time of *L* or *H*, *A* is the time of the candidate anchor, and *c* is a constant, i.e. the interval between segmental landmark *A* and tone *T* is *c* ms, or equivalently, tone *T* occurs *c* ms after segmental landmark *A*.

(3) $T\text{-}A = c$

Candidate anchors defined in terms of proportions of constituents were evaluated by fitting the models shown in (4), where *p* represents the proportion of the relevant constituent, and is estimated in fitting the models. These models state that the interval between *T* and the onset of the relevant constituent is a fixed proportion of the duration of that constituent. For example, the onset of the syllable is C1, so *T*-C1 is the interval between *T* and the onset of the syllable, and the offset of the syllable is Off, so Off-C1 is the duration of the syllable. So if *T*-C1 = 0.5×(Off-C1) then *T* is aligned to middle of the syllable.

(4)

Syllable:	$T\text{-C1} = p(\text{Off-C1})$
Rime:	$T\text{-V1} = p(\text{Off-V1})$
Interval:	$T\text{-V1} = p(\text{V2-V1})$

All models were fitted as linear mixed effects models using the *lmer* function from *lme4* (Bates et al 2014), with random effects by subject for the free parameter in each model, so each fitted model has two parameters: a fixed effect, and a corresponding random effect. The best anchors were taken to be those which yield the best fitting models of *L* and *H* respectively, assessed by comparing the deviances of the models (lower deviance indicates better fit to the data). The results are summarized in Table 1, where the deviances of the models for each candidate anchor are presented together with the standard deviation of the residual error, which is similar to the root-mean-squared-error of the fitted values[3].

[3] The difference is that the standard deviation of the residual error is estimated in fitting the model by MLE rather than being calculated from the residuals themselves.

Table 1
Deviances and standard deviations of the residuals of the models for each of the candidate segmental anchors for the onset, *L*, and offset, *H*, of the rising tone. The values for the best fitting model for each tone are in bold.

Tone:	*L*		*H*	
Anchor	deviance	s.d. of residual (ms)	deviance	s.d. of residual (ms)
onset of syllable (C1)	3841	106	4088	158
onset of vowel (V1)	3632	76	3949	127
offset of vowel	3473	59	3582	71
offset of syllable (Off)	3493	61	2981	27
onset of following vowel (V2)	3800	100	3267	43
proportion of syllable	**3201**	**38**	3001	29
proportion of rime	3240	41	3021	29
proportion of interval	3251	41	**2907**	**24**

The best fitting anchor for *L* is at 56% of the syllable duration, so *L* tends to occur a little over half way through the syllable. The best fitting anchor for *H* is 82% of the way through the vowel-to-vowel interval. This is generally a little after the syllable offset, but the interval-based characterization provides a better fit than aligning *H* to the end of the syllable because *H* tends to occur later when the onset of the following syllable is longer. These results are in line with previous studies of the Mandarin rising tone by Xu (1998) and Chen & Gussenhoven (2008), as is discussed in more detail below.

Although the tones remain close to these best anchor points, there are systematic deviations from the alignment targets as a function of syllable duration (Fig. 9) so tonal alignment is not invariant. The left panel shows *L* timing as a proportion of syllable duration, plotted against syllable duration. Since the mean best anchor for *L* is at 56% of syllable duration, this plot makes it easy to see the position of *L* relative to this anchor, plotted as a horizontal dashed line. It can be seen that *L* precedes this anchor at short syllable durations (i.e. fast speech rates), occurring as early as the syllable onset at the shortest durations. The pattern of timing of *L* identified by piecewise linear approximation is very similar. The *H* tone displays the converse pattern, following its anchor at short syllable durations, as shown in the right panel. Since the best anchor for *H* is a proportion of the vowel-to-vowel interval, *H* timing is plotted as a proportion of interval duration, with a dashed line at the mean anchor proportion, 0.82. *H* tends to occur much later than this anchor at the shortest durations, even occurring in the vowel of the following syllable. These plots also serve to illustrate that the speech rate manipulation was successful in eliciting a wide range of syllable durations.

The effect of syllable duration on deviation from these segmental anchors is statistically significant, as shown by fitting linear mixed effects model with the residuals of the strict segmental anchoring models as the dependent variable and syllable duration as predictor, with random slopes and intercepts by speaker, and comparing to models that eliminate the fixed effect of syllable duration with Likelihood Ratio Tests (*L* residuals: $\beta = 0.1$, $\chi^2(1) = 6.1$, $p < 0.05$; *H* residuals: $\beta = -0.05$, $\chi^2(1) = 7.6$, $p < 0.01$).

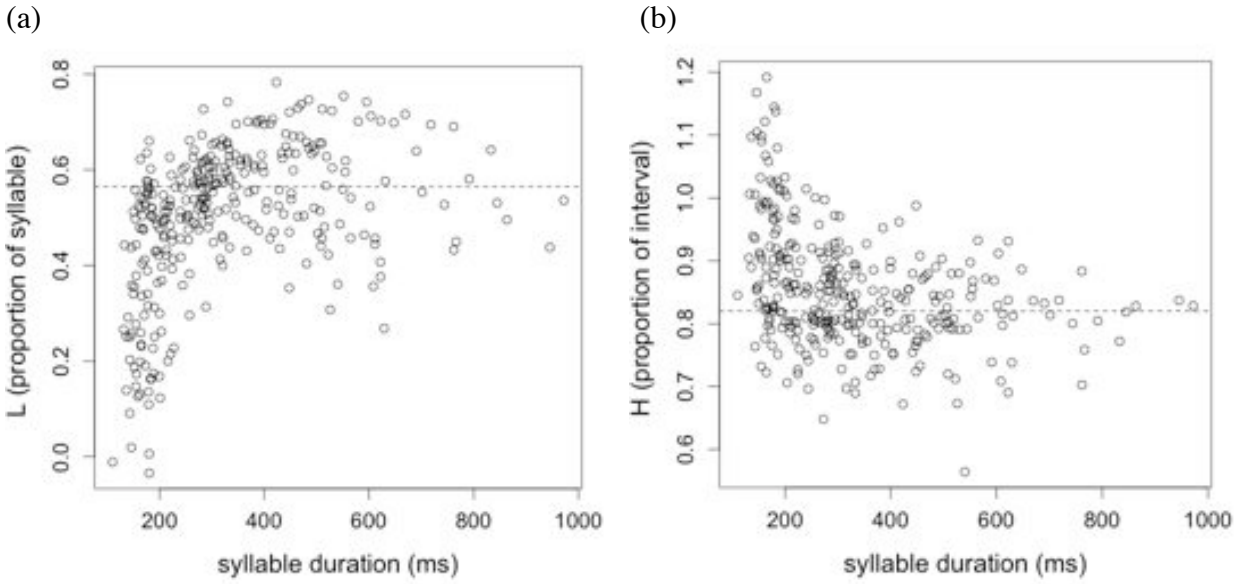

Fig. 9. Scatter plots of the timing of *L* and *H* as a function of syllable duration. *L* timing is plotted as a proportion of syllable duration (a) and *H* timing is plotted as a proportion of interval duration, with the timing of the segmental anchors indicated by dashed lines.

It should be noted that a purely linear dependence of residuals on duration could be captured by adding an intercept term to the segmental anchoring models, effectively adopting a tone-timing model of the form: *T*-C1 = *p*(Off-C1)+c[4]. Such a model could be interpreted as a segmental anchoring model, but the implied anchors would not constitute segmental landmarks in any obvious sense. For example, the implied segmental anchor for *L* would be 39 ms before the point 67% of the way through the syllable. In any case, we will see that the systematic deviations from strict segmental anchoring can be derived from the interaction of a simple segmental anchoring requirement with independently motivated constraints on tone realization, avoiding the need to stipulate such complex anchors.

The fact that at fast speech rates *L* precedes its anchor while *H* follows its anchor, means that the duration of the F_0 rise from *L* to *H* varies less than would be expected if these tones precisely tracked their respective anchors. As observed above, in the discussion of Fig. 2, this pattern of realization would be expected if speakers attempt to maintain a constant value for the magnitude of the rise, *M*,

[4] For example, according to the segmental anchoring model for L, the estimated timing of the low tone, $\hat{L}$, is given by the expression in (1).

(1) $\hat{L} - C1 = p(\text{Off} - C1)$

If the residuals of this model, i.e. the differences between the actual and estimated timing of L, $L - \hat{L}$, are a linear function of syllable duration, then the equation stated in (q) holds:

(2) $L - \hat{L} = b(\text{Off} - C1) + c$

Substituting (1) into (2) and rearranging yields (3), which has the same form as the segmental anchoring model with an intercept term, c, added, and a different coefficient for syllable duration (p+b in place of p).

(3) $L - C1 = (p + b)(\text{Off} - C1) + c$

and its slope, *S*, because that entails that the duration of the rise must remain constant too. However, the deviations from the segmental anchors are not large enough to maintain a constant rise duration, as illustrated in Fig. 9. This figure plots rise duration, *H-L*, against the duration of the interval between the segmental anchors for *L* and *H*. If the tones were consistently aligned to their anchors, then rise duration should be equal to the interval between anchors, and the points should cluster around the dashed '*y* = *x*' line. If rise duration were constant, the points should form a horizontal line. The observed pattern lies between these two extremes: rise duration varies substantially as a function of the interval between the anchors, but it falls short of longer intervals between anchors and does not decrease much as the interval between anchors falls below 200 ms, so rise duration can be substantially longer than the interval between anchors at short syllable durations. We will see in section 4 that this pattern can be analyzed as a compromise between the goals of aligning *L* and *H* to their segmental anchors and maintaining fixed targets for Magnitude and Slope (and hence for the duration of the rise).

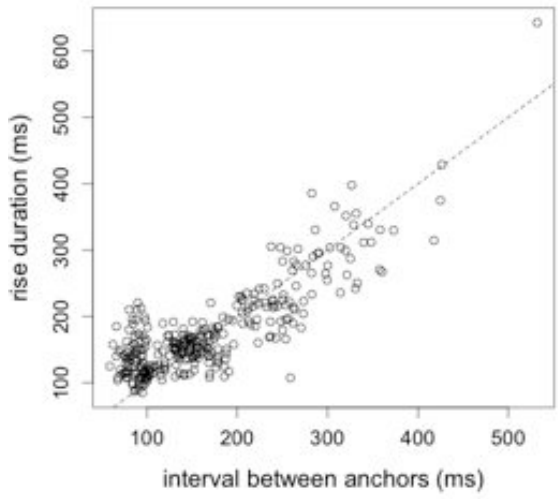

Fig. 9. Scatter plot of rise duration (*H-L*) as a function of the interval between the segmental anchors for *L* and *H*. The dotted line indicates where rise duration is equal to the interval between the anchors.

Similar patterns of variation in the alignment of *L* and *H* are reported by Xu (1998) in a related study of the effects of speech rate and syllable structure on the realization of Mandarin tones. His correlate of the onset of the rise, the acceleration maximum, occurs around 60% of the way through the estimated syllable duration for long syllable durations, but generally occurs steadily earlier as syllable duration decreases (pp. 198ff.). *H* occurs at, or after, the onset of the following vowel at the fastest speech rates, and occurs progressively earlier as syllable duration increases. In comparing the two studies it should be noted that, in addition to the different criteria for identifying *L*, the rising tones in Xu's study were also elicited in a different tonal context, between a high tone and a falling tone as opposed to the low_rising context employed here. A preceding high tone tends to push the onset of the rise later at short syllable durations due to the time it takes to realize the fall from the preceding high tone, as observed in the study by Chen & Gussenhoven (2008: 737, fig. 9), discussed next.

Chen & Gussenhoven (2008) investigated the timing of the onset of the rise in tone 2 under variation in syllable structure and degree of emphasis (simple statement, correction, and repetition of the correction in response to simulated mishearing), which elicited a wide range of syllable durations. In spite of the fact that the duration variation was elicited by varying emphasis rather than speech rate, the results for the timing of the rise onset are remarkably similar to those reported here: the timing of the onset of the rise is well approximated by a linear function of syllable duration with slope of 0.72 and intercept of -25 ms (measured from their Fig. 8, p.737), compared to the slope of 0.67 and intercept of -39 ms described above for the present data. That is, their *L* correlate occurs at about two thirds of the way through the syllable at the longest syllable durations, and occurs progressively earlier in shorter syllables. Some of the difference may be attributed to the fact that Chen & Gussenhoven's data includes rising tones preceded by high tones as well as low tones (see section 5.2 for further discussion of the effects of tone context).[5]

3.2. Magnitude and slope of the rise

Since rise duration increases as syllable duration increases, either the magnitude of the rise must increase as well, or the slope of the rise must decrease. That is, if the magnitude of the rise remains constant, then an increase in duration implies that the slope of the rise gets shallower (Fig. 2a). On the other hand, if the slope remains constant, then an increase in rise duration implies an increase in the magnitude of the rise (Fig. 2b). In fact we observe both effects: Magnitude increases with increasing rise duration (Fig. 10), but not sufficiently to maintain a constant slope, so slope decreases with increasing rise duration (Fig. 11).

The general pattern observed in Fig. 10 is that, as rise duration increases from its lowest values, *M* first increases rapidly then levels out. However, there is a set of outliers that do not conform to this pattern, having high rise magnitudes relative to their rise durations of around 100-175 ms. These higher values for *M* translate into high slopes relative to duration in Fig. 11. Most of these outliers are due to a single speaker using a higher, wider pitch range at normal speech rate (subject 2, plotted with open triangles). That is, this speaker used a higher, wider F_0 range throughout the sentences produced at normal rate, compared to her fast and slow renditions. We take these observations to indicate that *M* and *S* depend on both rise duration and pitch range. Our speakers mostly used pitch ranges of consistent and comparable sizes, so the observed variation in *M* and *S* is primarily a function of rise duration, and will be analyzed as such in the next section, but we will discuss how pitch range variation might be modeled in section 5.1.

[5] Chen & Gussenhoven only provide statistical analyses of *L* timing as a function of degree of emphasis rather than as a function of syllable duration, but it is apparent from their Fig. 8 that variation in *L* timing as a function of syllable duration is rather consistent across emphasis conditions – i.e. the effect of emphasis on tone timing is plausibly mediated by syllable duration.

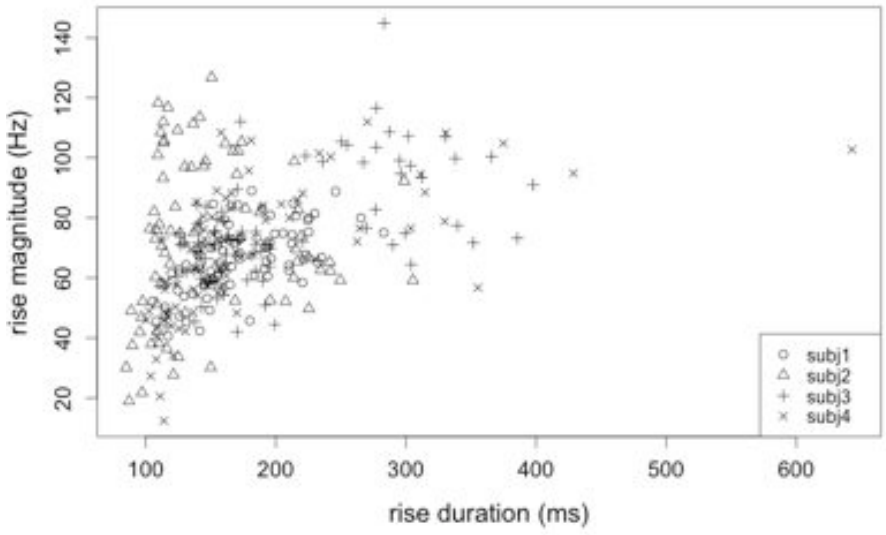

Fig. 10. Scatter plot of rise magnitude, *M*, as a function of rise duration, *H-L*. Speakers are differentiated by plotting symbols.

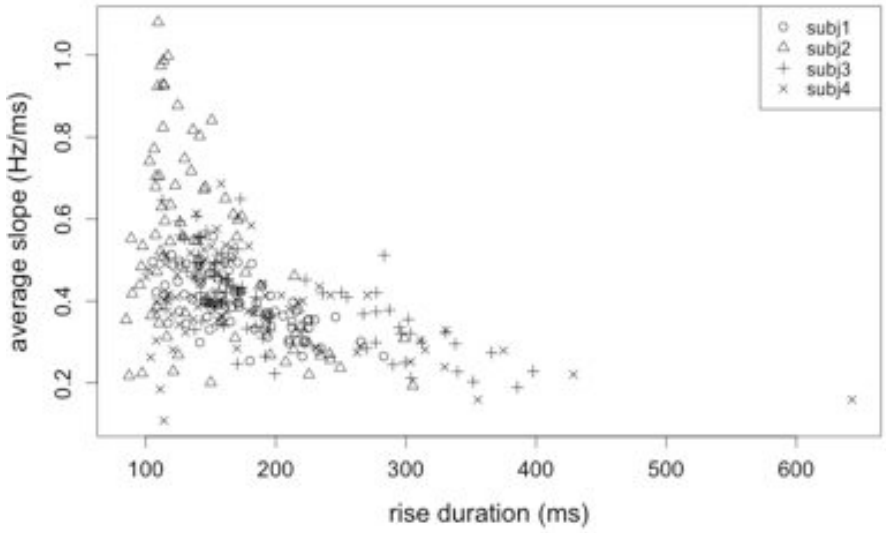

Fig. 11. Scatter plot of slope of the rise, *S*, as a function of rise duration, *H-L*. Speakers are differentiated by plotting symbols.

The effect of duration on magnitude is statistically significant: a linear mixed effects model predicting rise magnitude as a function of duration, with random slopes and intercepts by speaker, fits significantly better than a model according to which rise magnitude is constant, with the same random effects ($\chi^2(1)=8.8$, $p < 0.01$). A similar test shows that the effect of duration on slope is also significant ($\chi^2(1)=5.8$, $p < 0.05$). These results are unchanged if the extreme outlier with rise duration greater than 600 ms is excluded from the analyses.

The previous studies by Xu (1998) and Chen & Gussenhoven (2008) appear to describe somewhat different patterns with respect to rise magnitude and slope, although the data they report are not directly comparable in crucial respects, pertaining to the contexts in which the rising tones were elicited and the experimental manipulations that yield the duration variation. These differences are discussed in section 5, once we have a model in place that provides a basis for analyzing them.

3.3. Summary of the results

In summary, none of the properties of the rising tone are independent of speech rate. As speech rate increases, *L* shifts earlier relative to its anchor, *H* shifts later relative to its anchor, the magnitude of the rise, *M*, decreases, and the slope of the rise, *S*, increases. So we have no evidence for a division between specified and unspecified properties of the rising tone because such a division would lead us to expect some subset of the tonal properties to remain stable under variation in speech rate while the rest adjust to accommodate the invariant targets. In the next section we develop an analysis of the observed patterns of variation according to which the rising tone has specified targets for all of the properties studied here: *L*, *H*, *M* and *S*. It is not possible to realize targets for all four properties simultaneously, so the realization of the rising tone involves a compromise between the four targets. The nature of this compromise depends on segment durations, giving rise to rate-dependent variation in all four tone properties. This analysis is explained in more detail in the section 4.1, and is provided with a quantitative formulation in section 4.2.

4. Analysis

4.1. Outline of the analysis

The variation in the realization of the rising tone can be derived by positing that the rising tone has targets for all of the properties under consideration:

1. *L* should be aligned to a segmental anchor, A_L, which is about 60% of the way through the syllable.
2. *H* should be aligned to a segmental anchor, A_H, which is about 80% of the way through the interval.
3. The magnitude, *M*, of the rise should be T_M Hz.
4. The slope, *S*, of the rise should be T_S Hz/ms.

These targets conflict so it is not possible to realize them all. The nature of the conflict is illustrated in Fig. 12. A rise that meets the targets for magnitude and slope must have a fixed duration of T_M/T_S, since the slope is defined to be the magnitude of the rise divided by its duration, so satisfying these two targets would imply that the duration of the rise is invariant across speech rates and syllable structures.

But the duration between the segmental alignment targets for *L* and *H*, A_L and A_H, depends on segmental durations and thus varies with speech rate: the alignment targets are closer together at faster speech rates and farther apart at slower speech rates. Fig. 12 shows two speech rates, fast on the left and slow on the right with the segmental alignment targets closer together in the fast speech condition. The dashed triangles indicate the shape of an F_0 rise that satisfies the targets for *M* and *S*, i.e. it has a rise magnitude of T_M and a slope of T_S, and thus a duration of T_M/T_S. If the duration of the syllable and interval is such that the segmental alignment targets A_L and A_H are exactly T_M/T_S ms apart, then all of the targets can be realized, but if A_L and A_H are closer together (left of Fig. 12) or farther apart (right of Fig. 12) then satisfying the magnitude and slope targets implies failing to align one or both tones with their alignment targets.

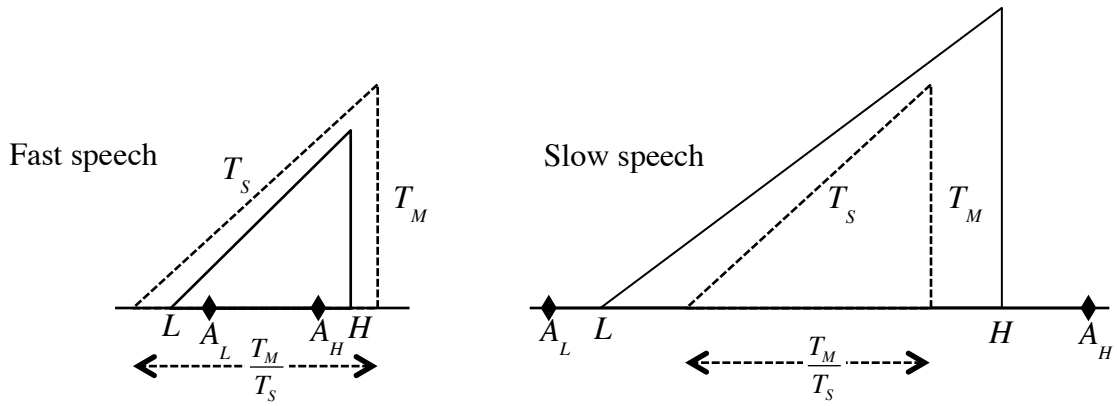

Fig. 12. Schematic illustration of the conflict between realizing the magnitude, slope and alignment targets for a rising tone. The dashed lines show the shape of the rise that satisfies the targets for rise magnitude and slope, while the solid lines schematize the actual slope and magnitude of the rise appropriate for the illustrated intervals between the alignments targets, A_L and A_H.

The observed pattern of tone realization as a function of speech rate follows if this conflict between the targets is resolved by compromise – that is, there is some deviation from each target rather than three of the targets being perfectly realized at the cost of large deviations from the remaining target. At faster speech rates, the duration between A_L and A_H is shorter than T_M/T_S (left of Fig. 12), so *L* precedes A_L and *H* follows A_H, bringing the rise duration closer to T_M/T_S, but it still falls short, so the rise is smaller than its target magnitude, T_M, although the slope is steeper than its target, T_S. Conversely, in slower speech, the duration between A_L and A_H is longer than T_M/T_S (right of Fig. 12), so *L* follows A_L and *H* precedes A_H, keeping the rise duration closer to T_M/T_S, but the rise is still too long, so the rise reaches a greater magnitude than its target, T_M, although its slope is a little shallower than its target, T_S. These are the qualitative patterns of variation in the rising tone as a function of segmental duration that we observed in the results above.

It is plausible that there should be targets for all of these properties because they all serve to distinguish tonal and intonational contrasts in Mandarin. Pitch levels and pitch slope are basic dimensions of tone perception in general (Gandour 1979, 1984) and in Mandarin in particular (Massaro, Cohen & Tseng 1985). Alignment of pitch events such as *L* and *H* also serves to distinguish Mandarin tones. Fig. 13 illustrates this point for the contrast between the high and rising tones (tones 1 and 2) following a low tone (tone 3), based on Xu (1997: Fig. 6). Following a low tone, it is necessary to produce a rising F_0 movement to reach the target for the level high tone, so in this context both high and rising tones are realized by rising F_0 contours. As observed by Chen & Gussenhoven (2008:743), the two tones differ substantially in the timing of the rise: in the high tone the rise begins near the onset of the syllable, but it begins nearer the middle of the syllable for the rising tone (as in the results reported above, also cf. Li 2003:64f.). At the other end of the rising trajectories, the F_0 peak for the high tone precedes the syllable offset, whereas it follows the syllable offset for the rising tone (also observed above). So the multiplicity of targets for the rising tone can be understood to reflect the multiplicity of cues to phonological contrasts: the rising tone is distinguished from others by cues on a variety of dimensions, and deviation from target values along any of those dimensions could reduce the distinctiveness of contrasts, and thus should be minimized. Possible bases for the specific target values observed here are discussed further in section 5.2, after we have obtained quantitative estimates of those target values based on the formalized analysis developed in the next section.

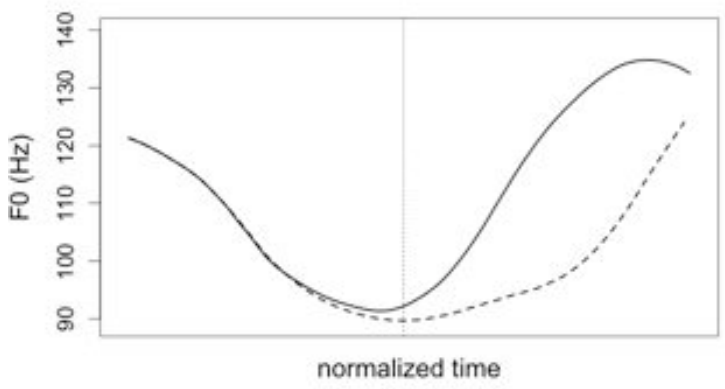

Fig. 13. F_0 trajectories of Low-High (solid) and Low-Rising (dashed) tone sequences produced on /mama/ segmental sequences by Mandarin speakers, based on Xu (1997: fig.6). The vertical dotted line marks the offset of the first vowel.

4.2. A constraint-based formalization of the analysis

So far the analysis of rate-dependent variation in the realization of the rising tone has been developed informally. In this section, we see that the proposal that the realization of the rising tone is a compromise between violable targets for segmental anchoring of the onset and offset of the rise and for slope and magnitude of the rise can be made quantitatively precise. The analysis depends on compromise between conflicting constraints, a concept that is central to the model of phonetic realization proposed by Flemming (2001), according to which phonetic grammars consist of weighted constraints and phonetic realizations are selected so as to minimally violate these constraints. Here we formalize the analysis of the Mandarin rising tone in these terms, showing that it accounts for the

qualitative patterns described above, and provides a reasonable quantitative fit to the data as well. The results provide evidence for the utility of this constraint-based approach to modeling phonetic realization.

The four targets for the rising tone are enforced by the constraints listed in Table 2. As discussed above, it is generally not possible to satisfy all of these constraints simultaneously, so the realization of the rising tone, given particular segmental durations, is selected to minimize violation of the constraints. The notion of minimal violation is defined by associating a cost of violation with each constraint. This cost is equal to the square of the deviation from the target (Table 2), and the total cost of a candidate set of values for the timing of *L* and *H* and the rise magnitude, *M*, is the weighted sum of its constraint violations, where the weights w_M, w_S, w_L, w_H, are positive numbers (5). The realization of the rising tone is then selected so as to minimize this total cost.

Table 2. Constraints and costs of violations.

Target	Constraint	Cost of violation
Magnitude	$M = T_M$	$w_M(M\text{-}T_M)^2$
Slope	$S = T_S$	$w_S(M/(H\text{-}L)\text{-}T_S)^2$
L alignment	$L = A_L$	$w_L(L\text{-}A_L)^2$
H alignment	$H = A_H$	$w_H(H\text{-}A_H)^2$

$$(5) \quad Cost = w_M(M - T_M)^2 + w_S\left(\frac{M}{H-L} - T_S\right)^2 + w_L(L - A_L)^2 + w_H(H - A_H)^2$$

Thus the parameters of the model are the target values and the constraint weights, and the outputs are values of *M*, *L* and *H*. Note that the constraints specify targets for *M*, *S*, and timing of *L* and *H*, but these quantities cannot be varied independently since the slope, *S*, is by definition equal to the magnitude of the rise, *M*, divided by its duration, *H-L*. Here we select values of *M*, *L*, and *H* and calculate *S* as *M*/(*H-L*), as shown in the formulation of the cost function for the Slope constraint (Table 2). In addition, the alignment targets, A_L and A_H, are not directly specified as parameters of the model, but are specified as proportions of the syllable and interval, respectively.

The cost function (5) is a convex function of *M*, *L*, and *H*, so it has a single minimum and solutions to the minimization problem can be found through gradient descent or other standard optimization algorithms[6]. The table in (6) shows an example of the evaluation of realizations of a rising tone given sample values of A_L and A_H together with the values of T_S, T_M and constraint weights estimated in the next section. The first four columns present candidate values for *L, H, S* and M with the target values for these properties given in the second row. The next four columns show the cost of violation of each constraint for each candidate tone realization, with the constraint weights in the second row. The final column shows the total cost incurred by each candidate. The minimum cost realization is in bold.

[6] Optimizations reported here were carried out using the Nelder-Mead algorithm as implemented in the R function `optim` (R core team 2013).

(6) Example of evaluation of realizations of a rising tone.

L (ms)	*H* (ms)	*S* (Hz/ms)	*M* (Hz)	*L* cost	*H* cost	*S* cost	*M* cost	Total cost
A_L= 140	A_H=268	T_S = 0.38	T_M = 76	w_L = 0.27	w_H = 0.56	w_S = 51500	w_M = 1	
122	**277**	**0.42**	**64**	**87**	**43**	**64**	**137**	**331**
140	268	0.38	49	0	0	0	749	749
140	268	0.59	76	0	0	2330	0	2330
140	340	0.38	76	0	2903	0	0	2851
68	268	0.38	76	1400	0	0	0	1400

Cost is minimized by a compromise between the conflicting constraints – that is, the optimal realization (the first candidate in (6)) involves modest violations of all of the constraints: *L* precedes its anchor, *H* occurs a little after its anchor, the rise is steeper than the target slope of 0.38 Hz/ms, and the magnitude target is undershot (the pattern schematized in the left panel of Fig. 12, above). The remaining candidate realizations in (6) illustrate the fact that these modest deviations from each target result in a lower total cost than a large deviation from one target together with perfect realization of the rest.

The optimality of compromise between conflicting constraints is a general consequence of minimizing the summed violations of constraints that penalize squared deviations from targets. The cost of violating a constraint grows rapidly as the magnitude of the deviation increases, so multiple small deviations from targets incur a lower cost than a single large deviation.

This is easiest to see by considering a conflict between two constraints. For example we can isolate the conflict between *L* ALIGNMENT and *H* ALIGNMENT by assuming a fixed rise duration. If the rise duration is greater than the interval between A_L and A_H, then placing *L* closer to A_L implies that *H* occurs later relative to A_H, and conversely, the closer *H* is to A_H, the earlier *L* must occur relative to A_L. This situation is illustrated in Fig. 14(a), which shows cost as a function of *L* timing. *L* ALIGNMENT requires *L* to be aligned to A_L while *H* ALIGNMENT requires *H* to be aligned to A_H. If the fixed rise duration is 200 ms, this second constraint implies that A_L should be aligned 200 ms before A_H. The cost of violating each constraint is proportional to the square of the deviation from the relevant target, plotted with thin lines in the figure. The sum of these costs is plotted with a heavy line. The optimal alignment of *L*, minimizing the summed cost, is a compromise between the two requirements (plotted as L_{opt} in Fig. 14). If the weights on *L* ALIGNMENT and *H* ALIGNMENT are equal then the optimum falls half way between A_H-200 ms and A_L, as illustrated in Fig. 14(a). If the weight on *H* ALIGNMENT is higher, then the optimum shifts closer to the target implied by that constraint, A_H-200 ms, as shown in Fig. 14(b), but the optimum falls between the two targets for all values of the constraint weights. That is, compromise is always optimal.

Not all cost functions derive compromise between conflicting constraints. For example, compromise would not follow if the cost of violating a constraint were equal to the absolute deviation from the target. As illustrated in Fig. 14(c), it is optimal to realize *L* at the higher-weighted target, which is A_H-200 ms in this example, with no compromise between the conflicting constraints[7].

[7] In the case of equal weights, this model would assign equal cost to all values of *L* between A_H-200 and A_L.

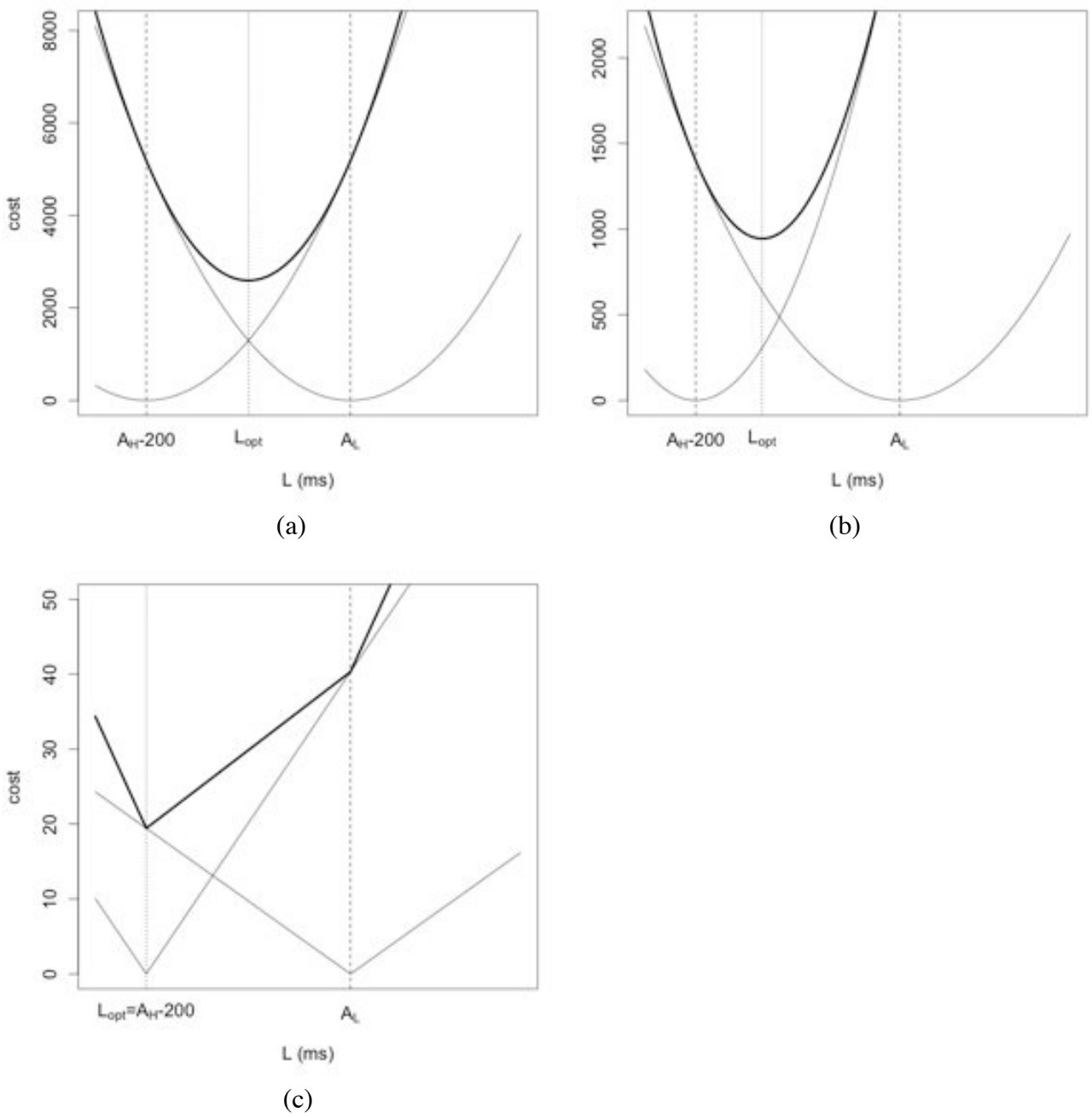

Fig. 14. Plots of cost as a function of timing of *L*. Thin curves show the cost assigned by the constraints *L* ALIGNMENT and *H* ALIGNMENT while the thick curves show the sum of these constraint violations. The dashed vertical lines indicate the targets set by the alignment constraints, A_L and A_H-200 ms, while the dotted vertical line marks the minimum-cost value of *L*, labeled L_{opt}. In (a) and (b), cost is proportional to squared deviation, with equal constraint weights in (a) and higher weight on *H* Alignment in (b). In (c), cost is proportional to absolute deviation, with higher weight on *H* Alignment.

4.3. Fitting the weighted-constraint model to the experimental data

To see how well the proposed model fits the experimental data, we need to estimate values for the targets T_S, T_M, A_L and A_H, and the constraint weights w_M, w_S, w_L and w_H. For the alignment targets, A_L and A_H, this involves re-estimating the location of the segmental anchors for *L* and *H*, in the context of a model where they are not the only factors determining tone timing, so the relevant parameters are the proportions of the syllable/interval where the segmental anchor is located. In addition, it is the ratios of the constraint weights rather than their absolute values that affect the predictions of the model, so one weight can be fixed at an arbitrary value. We set $w_M = 1$.

Estimating the model parameters is fairly complex because the model derives multiple variables (*M, L, H*), and each model parameter affects the predictions for all three variables, so the fit of the model has to be assessed with respect to all three variables simultaneously. In addition, there is no closed form solution to the problem of minimizing the cost function (5), so the predictions of the model given a set of parameter values have to be assessed through numerical optimization. The method we adopted was to estimate the model parameters through a hill-climbing search algorithm, maximizing the log-likelihood of the data given the model.

Starting from an initial guess as to the parameter values (targets and weights), the log-likelihood was calculated for each model that differs from the starting model by small changes in one or more parameters. The model with the highest log-likelihood was selected, then the process of checking neighboring models was repeated. The search continued until a maximum was reached – i.e. all neighboring models had lower log-likelihood than the current model. Step size for each parameter was set by preliminary investigation of the sensitivity of log-likelihood to changes in each parameter. After an optimum was reached, the step sizes were reduced and the search continued[8].

Given a set of model parameters, the predictions of that model were determined for each data point in turn by calculating A_L and A_H based on the observed segment durations and the current values of the alignment target parameters, then finding the values of *L*, *H* and *M* (and thus *S*) that minimize the cost function in (5) by numerical optimization, as described above. The goodness of fit of the model to the data was measured via an estimate of the log-likelihood of the data given the model, i.e. the log of the probability of the observed data given the model. To calculate the log-likelihood it is necessary to make assumptions about the probability of deviations from the model predictions. We assumed that errors in *L, H* and *M* were normally distributed, but not necessarily independent. The covariances between the errors *in L, H and M* were estimated from the errors in the fitted values.

The hill-climbing algorithm can get stuck in local maxima, so the procedure was repeated several times with different starting points. This procedure did identify two local maxima, and it is possible that neither is the global maximum, but the log-likelihoods and qualitative fits of the two solutions are very similar, so if there is a better solution it is likely that the resulting fitted values would be very similar to those reported here, and so would not make much difference to the assessments presented below since those depend on the overall fit to the data rather than on statistical comparisons of models or parameter estimates.

---

[8] Initial step sizes for each parameter were: A_L (proportion of syllable) 0.01, A_H (proportion of interval) 0.01, T_M 2, T_S 0.01, w_S 1000, w_H and w_L 0.01. For the second stage of the search, step size for T_M was reduced to 1 and step size for w_S was reduced to 500.

We report here on the fit of the model to a subset of the data that excludes the normal rate productions by speaker 2 because they were produced in a substantially higher and wider pitch range than the rest, as discussed above. Modeling this pitch range variation would in effect require allowing the magnitude target T_M to vary by speaker and by repetition of the materials, which would result in more parameters than we can estimate based on the available data, so for now we exclude this subset of the data, but approaches to modeling pitch range variation are discussed below, in section 5.1. We also excluded an extreme outlier with a rise duration of 643 ms, out of concern that this point could have excessive influence on parameter estimates (the next highest rise duration is 429 ms).

The optimal model parameters are as shown in (7). It is clear from some exploration of the likelihood surface that the confidence intervals on most of these parameter values are quite wide because a change in one parameter can often be offset to some extent by compensatory changes in other parameters. For example, the effects of an increase in T_M can be partially offset by decreases in T_S and w_S, accompanied by increases in w_L and w_H.

(7) Parameter values of the best-fitting model

	Target	Weight
Magnitude	$T_M = 76$ Hz	$w_M = 1$
Slope	$T_S = 0.38$ Hz/ms	$w_S = 51500$
L alignment	A_L is 58% of syllable duration	$w_L = 0.27$
H alignment	A_H is 82% of interval duration	$w_H = 0.56$

At this early stage of model development, it is most important to confirm that the proposed analysis of rising tone realization captures the qualitative patterns observed in the results of the experiment. That is, to verify that the informal analysis proposed in section 4.1 works as intended when explicitly formalized. The following plots and analyses demonstrate that this is the case. With respect to the timing of *L* and *H*, we observed that as segment durations decrease, *L* occurs progressively earlier than its anchor whereas *H* occurs progressively later than its anchor. This pattern is captured by the proposed model, as illustrated in Fig. 15. Fig. 15(a) shows predicted *L* timing plotted as a proportion of syllable duration against syllable duration together with the actual values of *L* plotted in gray, while Fig. 15(b) provides a similar plot of observed and predicted *H* timing as a proportion of interval duration (cf. Fig. 9, above). The estimates of the positions of A_L and A_H are barely changed from the positions estimated using the strict segmental anchoring models presented in section 3.1, but the addition of the slope and magnitude constraints accounts for systematic deviations of the tones from their anchors as a function of segmental durations: at short durations, *L* occurs early and *H* occurs late, relative to their respective anchors, to avoid excessive deviations from the slope and magnitude targets. So we can now see that it is not necessary to adopt more complex definitions of the segmental anchors to account for the observed patterns of *L* and *H* timing – all that is required is to take into account the interaction between tonal timing and the realization of targets for the slope and magnitude of the rise. The fact that *L* deviates from its anchor more than *H* follows from the lower weight on the *L* alignment constraint, compared to the *H* alignment constraint.

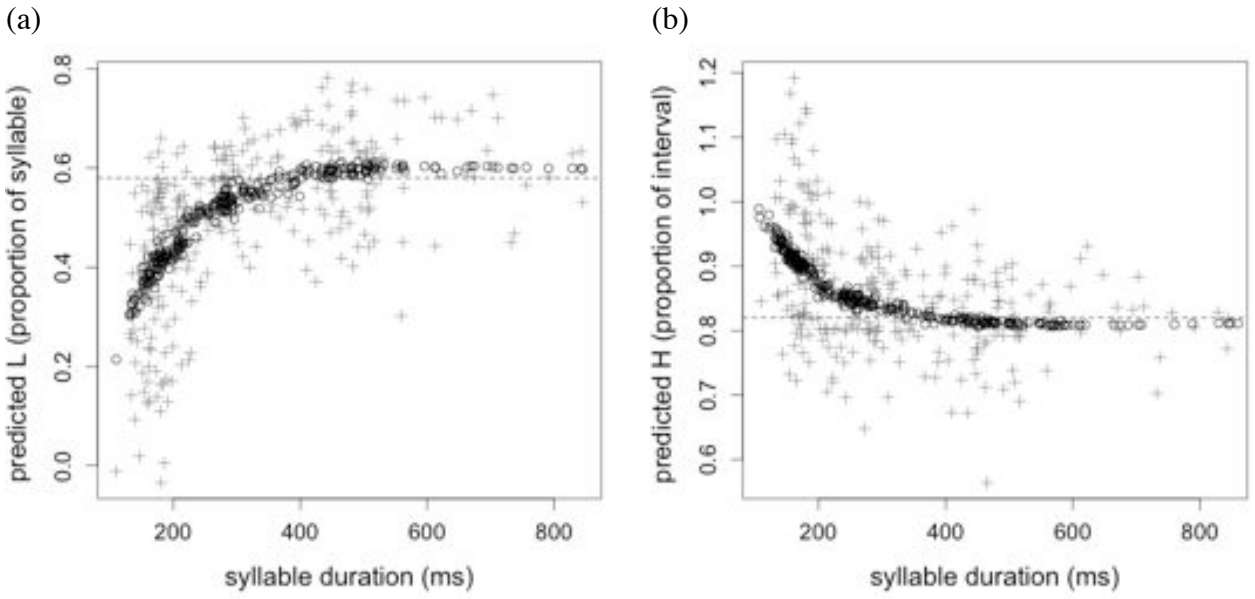

Fig. 15. Scatter plots of predicted *L* and *H* timing as a function of syllable duration (black circles), over actual *L* and *H* timing (gray crosses). (a) *L* timing is plotted as a proportion of syllable duration and (b) *H* timing is plotted as a proportion of interval duration, with the timing of the segmental anchors indicated by dashed lines.

The fitted values in Fig. 15 do not lie on a smooth curve because the model actually relates deviation of tones from their anchors to the duration between the alignment targets, and that duration depends on both syllable and interval durations whereas the plots in Fig. 15 only depict syllable duration in order to be comparable to the plots used to present the experimental results in Fig. 9, above. The predicted relationship between rise duration and the interval between A_L and A_H is shown in Fig. 16, together with the observed data, plotted with gray crosses.

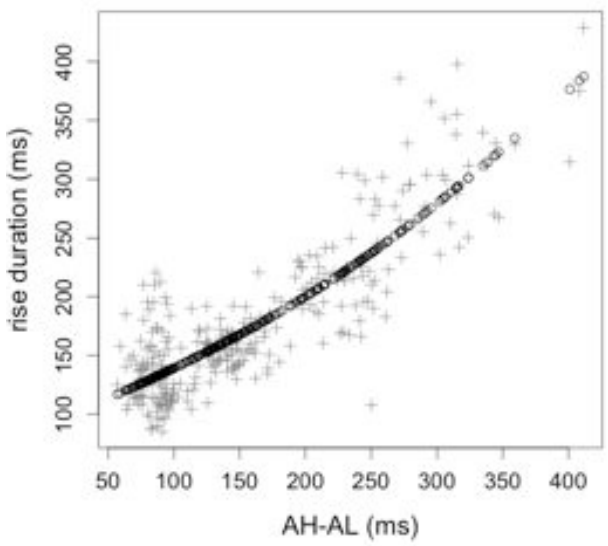

Fig. 16. Scatter plot of rise duration as a function of the interval between anchors, A_H-A_L (gray crosses) with values predicted by the model (black circles)

The other results we wish to derive are that, as the duration of the rise decreases, rise magnitude *M* decreases and slope *S* increases. The modeled relationship between *M* and rise duration, *H-L*, can be determined analytically from the partial derivative of the cost function with respect to *M*. The result is shown in (8), where *D* is the rise duration, *H-L*. The equation states that *M* is a weighted average of the target rise magnitude, T_M, and the rise magnitude that would result from a rise with the target slope value, T_S, and the observed rise duration, i.e. DT_S. The relative weights depend on the respective constraint weights and the square of the rise duration. The relationship between *S* and rise duration can then be derived by dividing the equation for *M* by *D*, since $S = M/D$ (9). These curves are plotted over the actual data in Fig. 17, where it can be seen that they capture the observed pattern: predicted *M* rises rapidly at short rise durations, then levels out, while predicted *S* declines as rise duration increases. Target rise magnitude, T_M, of 76 Hz is predicted to be achieved when rise duration is 200 ms, with undershoot at shorter durations and overshoot and longer durations. The converse pattern obtains for slope *S*. Note that the fitted values of *M* and *S* are not as good as these curves might suggest because the accuracy of the fitted values depend on the accuracy with which rise duration is modeled (see Fig. 18, below).

(8) $$M = \frac{w_M D^2 T_M + w_S D T_S}{w_M D^2 + w_S}$$

(9) $$S = \frac{w_M D^2 \frac{T_M}{D} + w_S T_S}{w_M D^2 + w_S}$$

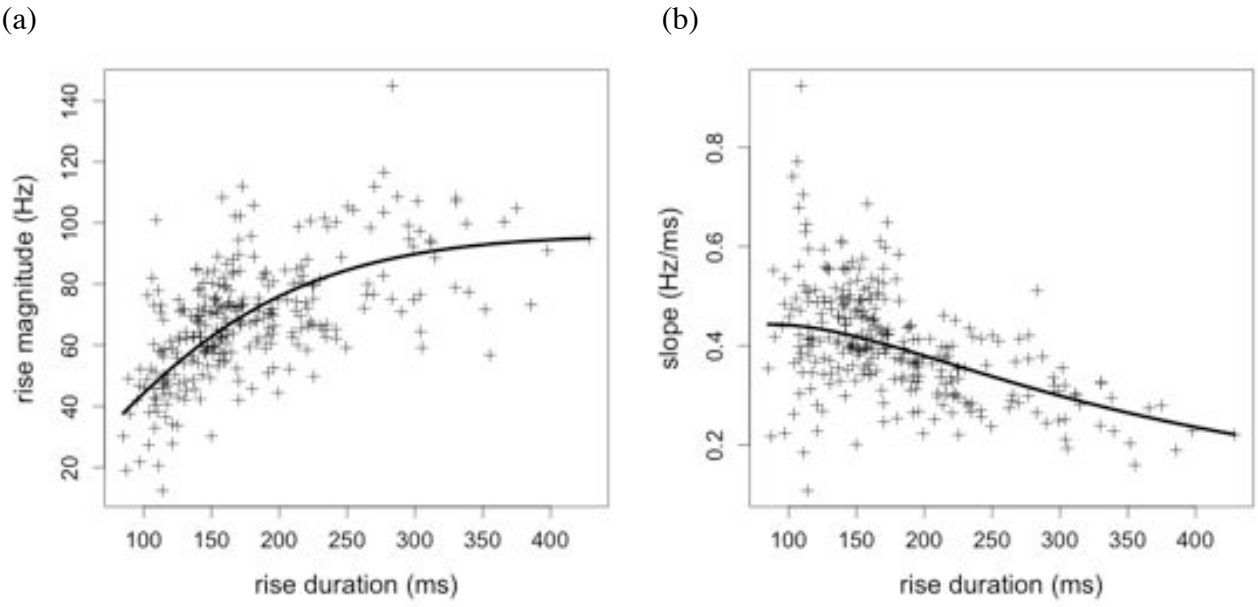

Fig. 17. Scatter plots of (a) rise magnitude, *M*, and (b) slope, *S*, as functions of rise duration, *H-L* (gray crosses). Black curves show the relationships derived from the model.

Although the model accounts for the key qualitative patterns observed in the experimental results, more detailed examination of the quantitative fit of the model to the data reveals areas in need of improvement. Model fit is summarized by the plots of observed against fitted values for *L*, *H*, *M* and *S* in Fig. 18.

Much of the variation in *H* is accounted for by the model (root-mean-square error (RMSE) is 23 ms), but *L* is modeled somewhat less accurately (RMSE = 37 ms). Some of this difference can be attributed to greater errors in the measurement of *L*, given that identifying the onset of the rise is much harder than identifying the F_0 maximum associated with *H* (as discussed in section 2.3), but the model also systematically overestimates the lowest values of *L*. That is, for observed values of *L* less than 50 ms, the fitted values are always higher than the observed values. Examining these cases, some are plausibly attributed to mismeasurement of *L*. For example, one speaker regularly produces the rising tone with a rising 'plateau' followed by a steeper final rise at the fast speech rate and, in these circumstances, the algorithm for identifying *L* is liable to place it at the beginning of the rising plateau when arguably it should be located later, at the onset of the steeper rise. However not all of the mismatches between predictions and observations can be explained in this way. The remaining modeling errors are not due to a failure to correctly model the relationship between syllable duration and *L* timing because other utterances produced with comparable syllable durations are produced with *L* occurring later than the model predictions. That is, substantial variance remains in the timing of *L* even after the effects of syllable duration are taken into account. The most promising account of these data is that there is variation in model parameters between, and perhaps within, speakers. For example, the fit to the *L* data can be significantly improved by allowing speaker-specific A_L alignment targets (ranging from 50% to 68% of syllable duration), while holding other parameters constant. However, more data from each speaker would be required to estimate these additional parameters with any confidence.

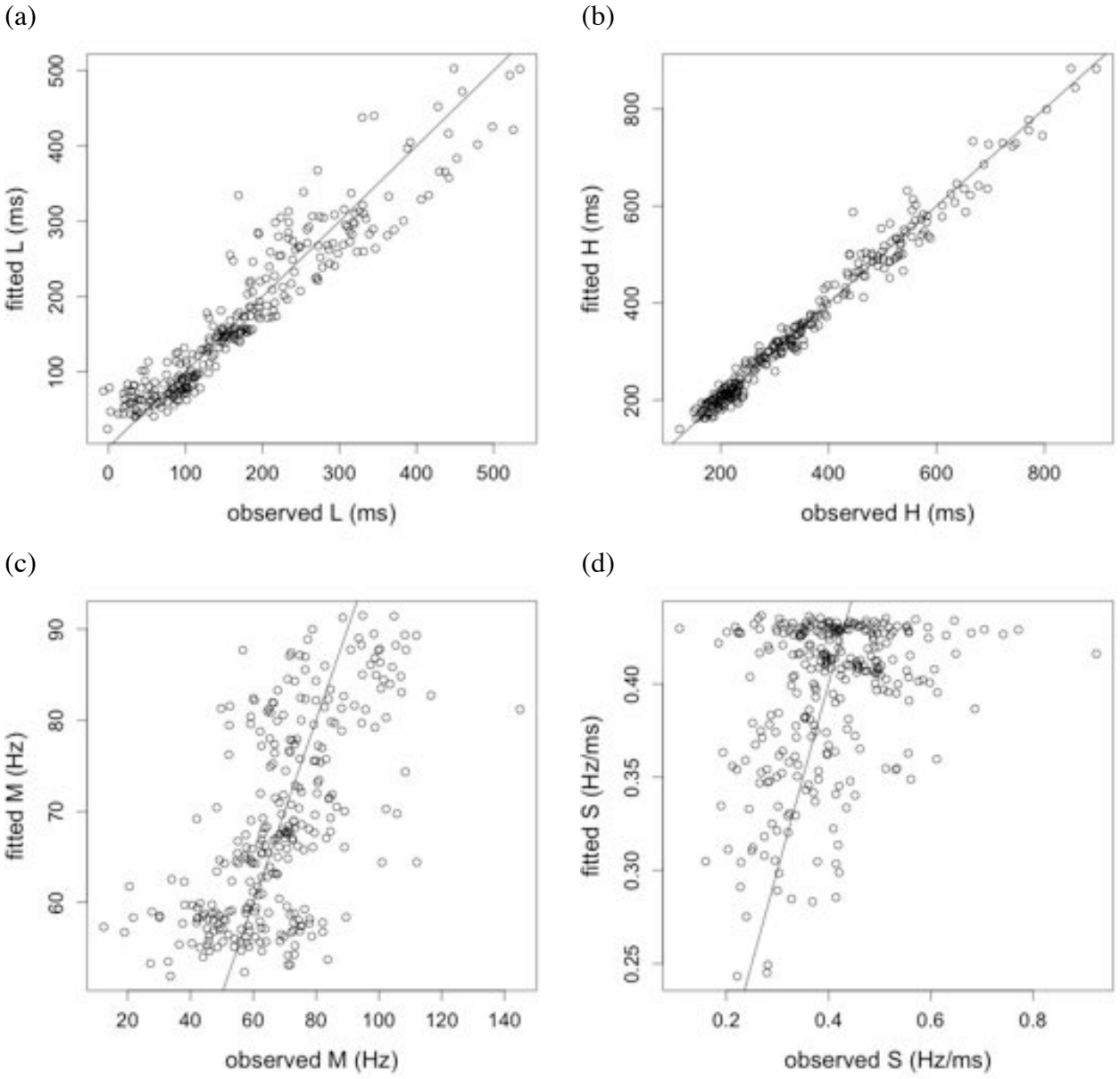

Fig. 18. Scatter plots of fitted against observed values for (a) *L*, (b) *H*, (c) *M* and (d) *S*.

Magnitude of the rise, *M*, and slope, *S*, are less accurately modeled (RMS errors of 15 Hz and 0.1 Hz/ms, respectively). A likely source of error here is that the model assumes a fixed value of T_M, when in actuality there is probably some variation between speakers, and even between repetitions of the word list by a single speaker. The fitted values of *S* are derived from the fitted values for *L*, *H* and *M*.

5. Developing the model of tone realization

We have seen in the previous section that variation in the realization of the rising tone as a function of segmental duration can be accounted for in terms of compromise between conflicting targets for the timing of the onset and offset of the rise, the slope of the rise, and the magnitude of the rise. In section 4.1 we hypothesized that the rising tone has targets for all of these properties because they are cues to tonal and prosodic distinctions. Now that we have estimates of the specific target values, we examine in more detail whether they are in fact interpretable in these terms. We will see that they are, but in some cases it is necessary to posit that the values currently posited actually derive from interaction between constraints, including an effort minimization constraint. These additions to the model are independently motivated by data from previous studies (particularly Xu (1998) and Gussenhoven & Chen (2008)), together with data from the present experiment on the realization of the rising tone when followed by a neutral tone.

5.1. The magnitude target - modeling variation in pitch range and prominence

The target for the magnitude of the rise, T_M, was estimated at 76 Hz above. While a rising tone obviously requires a positive target for rise magnitude, the specific value presumably varies as a function of pitch range and prosodic prominence. It is well known that speakers can vary the overall pitch height and span that they use from utterance to utterance (Ladd 2008:188ff.) and F_0 targets must vary accordingly. For example, in the model proposed by Pierrehumbert & Beckman (1988:175ff.), F_0 targets are assigned relative to reference lines indicating the top and bottom of the current pitch range. Selecting a pitch range then involves setting the levels of the high and low reference lines, and pitch targets for tones are specified in terms of proportions of the current pitch range, on a scale from 0 to 1. Applying such a model in the present context implies that T_M should be specified in terms of a proportion of the current pitch range, so when T_M is mapped onto a target in Hz it scales with pitch range. For example, a T_M of 0.5 of the difference between the high and low reference lines, would translate into a target rise of 50 Hz if the difference between the high and low reference lines is 100 Hz, but would yield a target magnitude of 90 Hz if the span of the pitch range is 180 Hz.

We did not systematically manipulate pitch range, so we do not have the data to test proposals concerning the nature of this variation, but, as noted above, one speaker did exhibit significant variation in pitch range between repetitions of the speech materials, producing the normal rate block with a higher, wider pitch range than the other two speech rates. Those data were excluded from the modeling in section 4 but can be accommodated by positing a different value for T_M, resulting from the expanded pitch range. The best fit to speaker 2's normal rate productions is obtained with $T_M = 91$ Hz if the other parameters are fixed at the values obtained above from fitting the model to the rest of the data.

T_M is also expected to depend on the prominence of a tone. For example a tone on a focused word has greater prominence than a tone on a non-focused word, and is thus associated with more extreme pitch targets. In Pierrehumbert & Beckman's model, the level assigned to a tone within the current pitch range depends on the prominence of the material with which that tone is associated – increasing prominence raises *H* and may lower *L*, which implies an increase in T_M. So a focused tone might have T_M equal to 0.8 of the current pitch range while a post-focal tone has a T_M of 0.5, for example. Given the same pitch range, the first target translates into a larger value for T_M in Hz than the second.

Chen & Gussenhoven (2008) provide evidence for this effect of prominence on the magnitude target, T_M. As discussed above, they elicited the rising tone under three levels of emphasis and found that rise magnitude increases with increasing level of emphasis (p.735). However it is not immediately obvious that this effect involves increasing T_M as a function of emphasis because syllable duration also increases with level of emphasis and we have seen above that rise magnitude increases with increasing duration even if prominence is unchanged. However, the behavior of the slope of the rise as the level of emphasis changes suggests that T_M does increase as prominence rises: In our data the slope of the rise declines steadily with increasing rise duration (Fig. 11), but slope does not decline steadily with increasing level of emphasis in Chen & Gussenhoven's data. Instead the slope of the rise is lowest for the non-emphatic words, increases under emphasis, but decreases slightly at the highest level of emphasis (p.739).

This pattern can be derived if increasing emphasis involves increasing both duration and T_M. As we saw above, for a fixed value of T_M, the proposed model predicts that slope should decrease with increasing duration (except for extremely short syllable durations), but, other things being equal, slope increases if T_M increases ((9) above). So whether an increase in emphasis results in an increase or decrease in slope depends on the balance between the contrary effects of increasing duration and of increasing T_M. Given this framework, we can interpret the low slopes in repeated, unemphasized words as resulting from a low value of T_M, reflecting low prominence. Slope is higher in the 'emphasis' condition because T_M is much higher, offsetting the reduction in slope that results from increased duration. Finally, slope decreases slightly in the 'more emphasis' condition because T_M is not much higher than in the 'emphasis' condition, and so fails to offset the reduction in slope that occurs with increasing duration. This analysis is consistent with the fact that the average rise magnitude is much lower in the non-emphatic condition than under either level of emphasis, but increased emphasis results in only a small increase in rise magnitude compared to the 'emphasis' condition (p. 735).

5.2. The slope target and tonal coarticulation

We have suggested that the rising tone has a slope target because a steep final rise distinguishes the rising tone from the other tones, which are either level or falling (with the exception of phrase-final low tones, as discussed above). But this line of reasoning implies that a steeper rise should be more distinct, whereas the model in section 4.3 posits a specific slope target of 0.39 Hz/s, and penalizes both steeper and shallower rises. A possible explanation for this slope target is that it is in fact a compromise between two conflicting constraints: a perceptually motivated Slope constraint which penalizes rises that fall below a minimum slope target, and a second constraint penalizing the articulatory difficulty of producing rapid F_0 changes, which favors shallower rises.

There are certainly physiological limits on rate of F_0 change (Xu & Sun 2002), but it is likely that transitions that approach those limits are also effortful to produce. Thus we can posit an effort minimization constraint that penalizes high rates of F_0 change, with cost effectively becoming infinite beyond the physiological maximum rate of change (cf. Lindblom 1983, Kochanski & Shih 2003). The precise form of the cost function remains to be established, but most of the effects described here only emerge at short syllable durations, suggesting that effort cost is minimal for transitions that fall below some threshold F_0 velocity, and rises rapidly for velocities above that threshold.

This effort constraint is independently motivated because it provides the basis for analyses of coarticulatory effects on the realization of the rising tone. Evidence concerning the nature of these effects comes from Chen & Gussenhoven (2008) where rising tones were elicited in the four contexts

obtained from all combinations of preceding high and low tones and following rising and falling tones, and from comparison of the present study on rising tones between a low tone and a rising tone with the results of Xu (1998), which studied rising tones produced between a high tone and a low tone.

5.2.1. Coarticulatory effects on the realization of the rising tone

An effort constraint penalizing rapid F_0 transitions creates conflicts between constraints on the realization of targets of adjacent tones, particularly where those targets differ substantially in F_0 level. For example, in our data the rising tones are preceded by low tones so only modest transitions are necessary between the two, but with a preceding high tone we have a $H_1.LH_2$ sequence with a high target at the end of the first syllable and a low target aligned near the middle of the second, so hitting the targets for H_1 and L would incur a high effort cost when syllable durations are short. The falling transition can be made shallower, reducing effort cost, by starting the fall early, lowering H_1, aligning L later than its alignment target, or undershooting L, realizing it above its pitch target (Fig. 19).

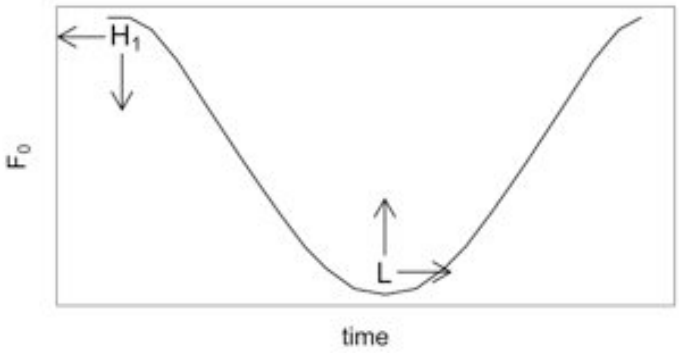

Fig. 19. Schematic F_0 contour for a high-rising tone sequence. Arrows indicate adjustments to H_1 and L that would reduce the slope of the initial falling transition.

Undershoot of L due to a preceding high tone is observed in Chen & Gussenhoven's (2008) data: when rising tones are produced in the unemphasized condition (i.e. with the shortest durations), the magnitude of the final rise is reduced when the tone is preceded by a high tone, compared to the context following a low tone (pp. 734f.). Averaged pitch tracks indicate that this is primarily due to undershoot of the L target of the rising tone (p. 733). This undershoot effect is also exhibited to a greater or lesser extent by all of the speakers in Xu (1998: Fig. 2).

Realizing L later in the syllable is not a viable solution because that would incur an offsetting cost by making the final rise too small and/or to steep. In fact the present study shows that, where the rising tone is preceded by a low tone, L is realized earlier in the syllable at short syllable durations, allowing more time for an adequate realization of the final rise. However we do find an effect of preceding context on timing of L in that Chen & Gussenhoven (2008:737f.) find that this tendency to retract L at short syllable durations is blocked when the preceding tone is high, presumably to avoid making the transition from the high tone too steep.

Chen & Gussenhoven find no evidence of adjustments to the preceding high tone in response to time pressure (pp.739f.). This is in line with the general observation that in Mandarin there is less deviation from tone targets at the end of syllables, compared to targets earlier in the syllable (Xu 1997), an effect

that would follow from higher weights on the constraints enforcing targets later in the syllable (Flemming 2011).

Note that the pattern of *L* undershoot following a high tone bears on a question that has been left open so far: Are there independent targets for the F_0 levels of the onset and offset of the rise, *L* and *H*, or is there a direct target for the magnitude of the rise? A direct target for rise magnitude would imply that *H* should be realized at a certain height above the onset of the rise, so if *L* is realized higher due to undershoot then we would expect a higher *H* peak as well. In fact Chen & Gussenhoven observe a smaller rise when *L* is undershot. This reduction in rise magnitude follows if the target F_0 level for *H* is independent of the actual level at which *L* is realized and thus is not raised in compensation for *L* undershoot.

The tendency to undershoot *L* at short durations when a rising tone is preceded by a high tone may also explain a final effect of preceding tonal context on the realization of the rising tone, involving the slope of the rise. The effect is revealed by a comparison of the present study, in which rising tones are preceded by a low tone, with Xu (1998), where the rising tones are preceded by high tones. In our study, average slope decreases steadily with increasing syllable duration (Fig. 11), but in Xu's study two out of four speakers show a different pattern where peak velocity is lowest at the shortest syllable durations then, as syllable duration increases, peak velocity first increases rapidly then falls. (One of the remaining speakers shows steadily decreasing velocity with increasing duration, while the last shows little change in peak velocity as a function of segmental duration). We hypothesize that this reduction in peak velocity at the shortest syllable durations is a side-effect of *L* undershoot conditioned by the preceding high tone. As just discussed, undershoot of *L* reduces the distance to the *H* target, which is equivalent to reducing T_M in terms of the current formulation of the tone model. As we have seen, the result of a smaller value for T_M is a shallower rise for a given rise duration ((8), (9) above). This reduction in rise slope only arises at the shortest durations because *L* undershoot only arises at the shortest syllable durations. Over the rest of the duration range, the relationship between slope and duration is similar whether the preceding tone is high or low.

This analysis raises the question why the reduction in slope at very short durations is only observed for two of the four speakers in Xu (1998). The most likely explanation is that the speakers have different constraint weights. The other two speakers keep the duration of the rise from shortening as much by pushing *L* nearer to the syllable onset and/or realizing *H* in the following vowel at short syllable durations (Xu's figures 3 and 5). This pattern would follow if these speakers place a higher weight on the Slope constraint relative to the *L* and *H* alignment constraints. Whether these differences in tone alignments are accompanied by differences in pitch levels is not clear since no relevant measurements are reported.

5.3. Alignment targets A_L and A_H

The alignment target for the onset of the rise, A_L, is estimated to be just over half way through the syllable. This target plausibly serves to differentiate the rising tone from contrasting high and low tones. As discussed in section 4.1, delaying the onset of the rise serves to distinguish the rising tone from the high tone, particularly in contexts where the offset of the preceding tone is low so the high tone is realized with a rising F_0 movement beginning at the onset of the syllable. The timing of the onset of a final rise has also been shown to be a cue to the distinction between low (tone 3) and rising tones, with a later rise onset cuing a low tone (Shen, Lin & Yan 1993, Moore & Jongman 1997). So a

syllable-medial rise onset distinguishes the rising tone from both the high and low tones. This is particularly clear for tones in pre-pausal contexts where tone 3 is realized with a final rise (Fig. 3), but even in medial contexts, where this final rise is generally absent, a later rise onset implies a greater duration of low F_0, which is likely to cue a low tone.

The alignment target for the peak of the rise, A_H, is estimated to be about 80% of the way through the associated vowel-to-vowel interval. We suggest that this apparent target is shaped as much by the need to keep the *H* peak from intruding on the realization of the following tone as it is by any cue-related target of the rising tone. That is, the estimated target is derived from analysis of rising tones that are followed by a second rising tone, so the *H* target of the first tone is followed by an *L* target in the next syllable. This means that if the *H* peak is realized later, it is closer it is to the *L* target, requiring a faster transition between the tones, which incurs a greater violation of the Effort constraint. So the targets for the level and alignment of *L* in conjunction with the Effort constraint favor early realization of the *H* peak. The observed timing of *H* can then be analyzed as a compromise between these constraints and a constraint preferring late alignment of the *H* peak, perhaps at the end of the interval.

Evidence for this line of analysis comes from data showing that the *H* peak is aligned later when the following tone is neutral (Li 2003, Chen & Xu 2006). This pattern follows if intrusion of the rising tone onto a following neutral tone incurs a lower cost than disruption of a full rising tone, which is consistent with the proposal that the neutral tone has a weak target (Chen & Xu 2006:67).

The neutral tone is the only tone that can appear on short, 'weak' or unstressed syllables. The F_0 pattern of the neutral tone has been described as varying primarily as a function of the previous tone (Lee & Zee 2014). In the context elicited here, between a rising tone and a falling tone, the neutral tone is generally realized with a falling F_0 contour, sometimes followed by a rise towards the high onset of the following falling tone. In our data, this rise is most often observed with the syllable [mən] (in [ɻə̆nmən] and [tsə̆nmən]), the only syllable with a coda nasal, and is generally confined to the final nasal (as in Fig. 20(a)), suggesting that the low target for the neutral tone may be aligned near the end of the vowel in both open and closed syllables.

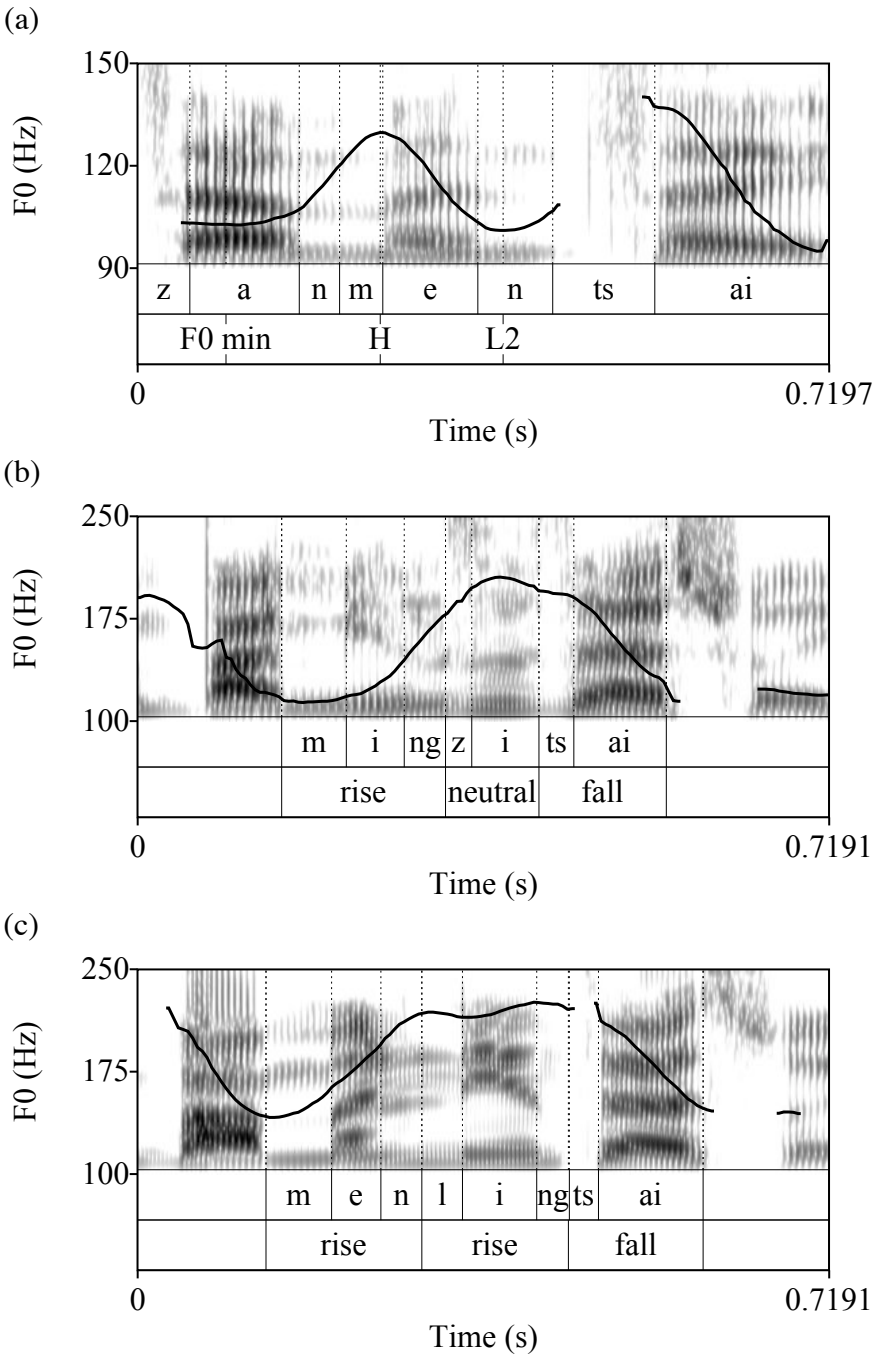

Fig. 20. Pitch tracks and spectrograms of bisyllabic words illustrating (a) an unreduced realization of a neutral tone following a rising tone in [tsănmən] 'we', (b) a reduced neutral tone following a rising tone in [mĭŋtsɨ] 'name' and (c) a reduced rising tone in [mənl̆iŋ] 'doorbell'. Each word is followed by a falling tone, on the word [tsai]. The figures are plotted with the same time scale.

Given that the neutral tone has a low target, the transition from a rising tone to either a neutral tone or a second rising tone involves a falling F_0 movement (compare Fig. 20(a) to Fig. 4). However, there are a number of differences between the realizations of the two tone sequences. The first, mentioned above, is that this falling transition tends to start later with a following neutral tone than with a following rising tone. Li (2003) and Chen & Xu (2006) find that the *H* peak of a rising tone followed by a neutral tone regularly occurs in the vowel of the neutral-toned syllable although it precedes the vowel of a syllable bearing a low or rising tone at similar speech rates. We observe a similar pattern in our data: for a given interval duration, *H* tends to occur later when the following tone is neutral than when the following tone is rising (Fig. 21).

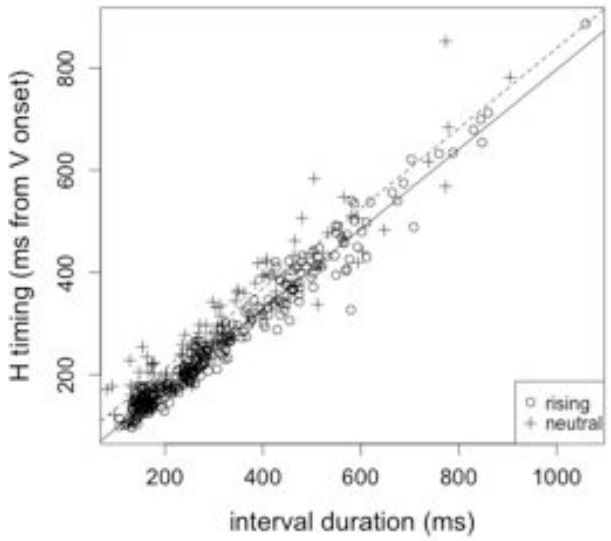

Fig. 21. Timing of *H*, relative to vowel onset, plotted against interval duration for tones preceding rising tones (open circles) and neutral tones (crosses). Least-squares linear regression lines are plotted for each context (a solid line for the context before a rising tone, and a dashed line for the context before a neutral tone).

We do not have sufficient data from the neutral tone context to fit a full constraint-based model, but we can provide a simple test of the effect of following context on *H* timing by using interval duration to control for segmental duration. The timing of the *H* peak relative to vowel onset is modeled quite accurately by a linear function of interval duration, as illustrated in Fig. 21. So we tested for effects of context on *H* timing by fitting a linear mixed effects model predicting timing of *H* relative to vowel onset from interval duration, context and the interaction of the two, with random intercepts and slopes by speaker corresponding to all fixed effects. The analysis shows a significant effect of context ($Z = 3.3$, $p < 0.001$), with *H* occurring about 40 ms later preceding a neutral tone. No difference could be detected between the slopes of the relationships between *H* timing and interval duration in the two contexts (i.e. the interaction between interval duration and context is not significant, $\beta = 0.002$, $Z = 0.06$). The two outliers above the neutral tone regression line both involve the word [pjǎnji] 'cheap', where measurement of interval duration is uncertain because the second syllable was generally realized

without any obvious distinction between glide and vowel. If this item is excluded, the difference in *H* timing between contexts is reduced to 30 ms, but remains significant ($Z = 2.1, p < 0.05$).

We suggest that this difference in the timing of *H* results from a difference in the weights on the constraints enforcing the pitch level targets of the rising and neutral tones, respectively. As described above, realization of the low target of a rising or neutral tone conflicts with the realization of a preceding *H* target, due to the Effort constraint on the transition between the two. The optimal compromise at most speech rates is to undershoot the target for the level of the low tone and to align *H* earlier than its target, making the transition from *H* to *L* slower, but the balance between *L* undershoot and *H* misalignment differs depending on whether the following tone is neutral or low. This is because the constraint enforcing the target level for the neutral tone carries a lower weight than the constraints enforcing the low targets for full tones (cf. Chen & Xu 2006:67), so with a following neutral tone the optimal compromise places *H* closer to its alignment target near the end of its interval, at the cost of greater undershoot of the *L* target. The lower weight on the pitch target of the neutral tone may be related to the fact that neutral tones are adequately differentiated from full tones by their short duration, so a precise F_0 contour is less important for their identification (cf. Li 2003).

The predicted difference in the amount of undershoot of *L* targets in rising and neutral tones is apparent from an examination of utterances produced at fast speech rate. At short syllable durations the neutral tone is often undershot to the extent that it is realized by a high plateau in the trajectory from the preceding rising tone to the following falling tone (e.g. Fig. 20(b)), and in many cases not even a plateau is apparent. Extreme undershoot of the *L* target of the second rising tone in a target word is also observed at fast speech rates, but it is more limited than undershoot of the neutral tone at equivalent speech rates (e.g. Fig. 20(c)). There is a local minimum in the F_0 trajectory of the neutral tone in only two out of 31 utterances in the fast speech rate condition, but half of the 106 second-syllable rising tones are realized with an F_0 minimum in the same rate condition. This difference is partially attributable to the fact that neutral-toned syllables are shorter than full-toned syllables, but the alignment target for *L* in the rising tone is only about half way through the syllable, whereas the low target for the neutral tone is near the end of the vowel, so the interval from *H* to *L* is comparable in the two tone sequences.

Having clarified that the underlying *H* alignment target is probably close to the end of the interval, we are still left with the question whether this target can be interpreted as realizing a cue to the identity of the rising tone. It is possible that late alignment of *H* distinguishes the rising tone from a high tone, at least in contexts where the following tone is low: Averaged pitch tracks presented in Xu (1997: Fig. 3) indicate that, preceding a tone with a low onset, the high tone peaks before syllable offset whereas the rising tone peaks after syllable offset. However, a more general characterization of this difference is to say that F_0 is rising more steeply at syllable offset in the rising tone, regardless of context – other tones are generally level or falling at syllable offset. If F_0 is rising at syllable offset then the peak has to occur after syllable offset since it takes time to change the direction of the F_0 trajectory, and the faster F_0 is rising, the later the peak is likely to occur. This line of reasoning suggests the possibility that there is actually a target for a rising slope at syllable offset, and that the late alignment of the *H* peak is a side effect of realizing this target (Xu & Wang 2001).

6. Conclusions

In this study we have explored the nature of the phonetic targets for the Mandarin rising tone by examining the realization of this tone across a range of syllable durations. The rationale behind this approach is that tonal properties that have specified targets should not vary with syllable duration, whereas properties that are not governed by target specifications should vary to accommodate the realization of specified targets. However, we found that all of the properties under examination varied systematically as a function of segmental duration: As syllable durations shorten, the onset of the rise occurs earlier in the syllable and the offset occurs later, the magnitude of the rise decreases, and the slope of the rise increases.

We have seen that these patterns of variation can be derived from an analysis according to which the rising tone has targets for all of these properties. These targets cannot all be realized since they conflict, and the conflict is resolved by compromise between the targets. That is, modest deviations from each target are preferred over a large deviation from a single target. This analysis was formalized in a framework where targets are enforced by violable, weighted constraints and phonetic realizations are selected so as to minimize the summed violations of the constraints. The resulting model derives the qualitative patterns of tone realization observed in the experimental data, and provides a reasonable quantitative fit to the data also.

These results carry implications for theories of tonal implementation in particular, and for theories of phonetic realization in general. With regard to tonal implementation, the results show that contour tones can have targets that pertain to both the endpoints of a pitch movement and to properties of the transition between those endpoints, such as the slope of the transition. So a model of tonal realization based purely on point targets and general interpolation mechanisms is insufficient. The more general implication is that tones (and presumably segments) can be overspecified in the sense of having mutually incompatible targets. This implies that phonetic realization incorporates a mechanism for resolving conflicts between targets – we have adopted a mechanism based on minimizing the summed violations of weighted constraints. This approach to phonetic realization is motivated by Flemming (2001) based primarily on analyses of coarticulatory phenomena. The present case is interestingly different in that the constraint conflict is inherent to the targets of a single tone, whereas coarticulation involves conflict between the demands of targets for neighboring segments or tones and an effort-minimization constraint on the transition between targets (cf. section 5.2.1). We close by briefly considering likely sources of inherent conflict between targets, and thus where additional examples of overspecification might be found.

We have hypothesized that the range of targets for the Mandarin rising tone reflect the fact that there are multiple cues that distinguish tones from each other or mark prosodic distinctions, such as prominence, through tone realization. That is, there are constraints enforcing targets corresponding to each cue. If this hypothesis is correct then we should expect most tones and segments to have multiple targets because contrasts are normally realized by multiple cues. However these targets need not be incompatible. For example, preceding vowel duration and Voice Onset Time are both cues to voicing in stops, but there is no inherent incompatibility in realizing any particular combination of values of these two properties. The targets of the rising tone conflict because they involve properties of a single entity, a rising F_0 movement, which are related by definition – e.g. slope is by definition equal to rise magnitude divided by rise duration. Diphthongs could constitute a comparable case of overspecification because they involve formant movements whose salient characteristics include onset and offset frequencies, slope and duration. The canonical F2 trajectory in the English diphthongs /aɪ,

aʊ, ɔɪ/ is comparable to the F_0 trajectory of the rising tone, consisting of an initial F2 plateau, followed by a rise (/aɪ, ɔɪ/) or fall (/aʊ/) to the offset of the vowel. There is also evidence for comparable effects of speech rate on the realization of these F2 trajectories: as speech rate increases, the onset of the F2 movement occurs proportionately earlier in the vowel (Gay 1968, Weismer & Berry 2003), and for many speakers the magnitude of the F2 movement tends to decrease, while its peak velocity increases (Dolan & Mimori 1986, Weismer & Berry 2003). (Gay (1968) observed a decrease in the magnitude of the rise but no effect on its velocity). These patterns are consistent with targets for the duration, magnitude, and slope of the F2 rise, all of which are plausible cues to the quality of the diphthong (e.g. Bond 1978, Morrison & Nearey 2007, Nábělek, Czyzewski & Crowley 1994).

Another possible source of conflict between targets is physical incompatibility. For example, the maximum F2 that a speaker can produce decreases as F1 increases, so it is not possible to produce a vowel with F1 as in low [a] but F2 as in high front [i]. This is because a high F2 requires a narrow constriction in the palatal region, which necessarily results in a relatively low F1 (e.g. Fant 1970, ch. 1.4). Consequently simultaneous targets for high F1 and F2 would conflict.

Positing conflicting targets of this kind might provide the basis for an analysis of variation in the realization of the 'tense' variant of the low front vowel /æ/ found in particular phonological (and lexical) contexts in a number of dialects of North American English (e.g. Labov et al 2006:173ff.). In at least some Mid-Atlantic dialects, tense /æ/ varies between a raised monophthong, close to [ɛ], and a diphthong that could be transcribed as [eæ], in which F1 rises and F2 falls. This variation might be analyzed as deriving from different compromises between incompatibly high F1 and F2 targets, perhaps depending on vowel duration. The monophthong realization then represents a compromise in which a physically possible combination of F1 and F2 values is achieved by allowing both F1 and F2 to fall short of their high targets. Given sufficient duration, the conflict is instead resolved by sequencing the incompatible targets, high F2 first, then high F1, resulting in a diphthong.

Both examples are speculative at this point, but serve to illustrate the potential for overspecification analyses. Further research is required to establish the generality of this phenomenon.

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.

Bates, Douglas; Maechler, Martin; Bolker, Ben and Walker, Steven (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-5. http://CRAN.R-project.org/package=lme4

Boersma, Paul & Weenink, David (2009). Praat: doing phonetics by computer [Computer program]. Version 5.1.12. Retrieved from http://www.praat.org/

Bond, Z.S. (1978). The effects of varying glide durations on diphthong identification. *Language and Speech* 21, 253-263.

Caspers, J. & van Heuven, V. (1993). Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. *Phonetica* 50, 161-171.

Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics* 36, 724-746.

Chen, Yiya, & Xu, Yi. (2006). Production of weak elements in speech – evidence from F_0 patterns of the neutral tone in standard Chinese. *Phonetica* 63, 47-75.

D'Imperio, Mariapaola. (2000). *The Role of Perception in Defining Tonal Targets and Their Alignment*. PhD dissertation, The Ohio State University.

del Giudice, Alex, Ryan Shosted, Kathryn Davidson, Mohammad Salihie, & Amalia Arvaniti (2007). Comparing methods for locating pitch "elbows". *Proceedings of the 16th International Congress of Phonetic Sciences*, 1117-1120.

Dolan, William B., & Mimori, Yoko (1986). Rate-depedent variability in English and Japanese complex F2 transitions. *UCLA Working Papers in Phonetics* 63, 125-153.

Farnetani, Edda & Kori, Shiro (1986). Effects of syllables and word structure on segmental durations in spoken Italian. *Speech Communication* 5, 17-34.

Flemming, E. (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18, 7-44.

Flemming, E. (2011). La grammaire de la coarticulation. Mohamed Embarki and Christelle Dodane (eds.) *La Coarticulation: Des indices à la Représentation*, L'Harmattan, Paris, 189-211.

Gandour, Jack (1984). Tone dissimilarity judgments by Chinese listeners. *Journal of Chinese Linguistics* 12, 235-261.

Gandour, Jack. (1979). Perceptual dimensions of tone: Thai. In: South-east Asian Linguistic Studies Vol. 3, edited by Nguyen Dang Liem. 3: 277-300. Pacific Linguistics, the Australian National University.

Goldsmith, J.A. (1976). *Autosegmental Phonology*. Ph.D. thesis, MIT.

't Hart, Johan, René Collier & Antonie Cohen (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Kochanski, G., & Shih, C. (2003). Prosody modeling with soft templates, *Speech Communication* 39(3-4), 311-352.

Kochanski, G., Shih, C., & Jing, H. (2003). Quantitative measurement of prosodic strength in Mandarin. *Speech Communication* 41, 625-645. http://dx.doi.org/10.1016/S0167-6393(03)00100-6.

Labov, William, Ash, Sharon, & Boberg, Charles (2006). *The Atlas of North American English*. Mouton de Gruyter, Berlin.

Ladd, D.R. (2004). Segmental anchoring of pitch movements: autosegmental phonology or speech production? In: Quene, H., van Heuven, V. (eds), *On Speech and Language: Essays for Sieb B. Nooteboom*, Utrecht: LUT, 123–131.

Ladd, Robert D. (2008) *Intonational Phonology*, 2nd Edition. Cambridge University Press, Cambridge.

Lee, Wai-Sum, and Zee, Eric (2014). Chinese phonetics. C-.T. James Huang, Y.-H. Audrey Li and Andrew Simpson

Li, Zhiqiang (2003). *The Phonetics and Phonology of Tone Mapping in a Constraint-Based Approach*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Lindblom, Björn (1983). Economy of speech gestures. Peter MacNeilage (ed.) *Speech Production*. Springer Verlag, New York. 217-246.

Lin, Hwei-Bing, & Repp, Bruno H. (1989). Cues to the perception of Taiwanese tones. *Language and Speech* 32:1, 25–44.

Massaro, D.W., Cohen, M.M., & Tseng, C. Y. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics, 13,* 267-289.

Moore, Corinne B., and Jongman, Allard (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America* 102, 1864-1877.

Morrison, Geoffrey Stewart, and Nearey, Terrance M. (2007). Testing theories of vowel inherent spectral change. *Journal of the Acoustical Society of America* 122, EL15-EL22.

Nábělek, Anna K., Czyzewski, Zbigniew, & Crowley, Hilary (1994). Cues for the perception of the diphthong /ɑɪ/ in either noise or reverberation. Part I. Duration of the transition. *Journal of the Acoustical Society of America* 95, 2681-2693.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Ph.D. thesis, MIT.

Pierrehumbert, Janet, & Beckman, Mary (1988). *Japanese Tone Structure*. Linguistics Inquiry Monograph 15, MIT Press, Cambridge.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramsay, J., & Ripley, B. (2013). pspline: Penalized Smoothing Splines. R package version 1.0-16. http://CRAN.R-project.org/package=pspline

Shen, Xiaonan Susan, Lin, Mascan & Yan, Jingzhu (1993). F0 turning point as an F0 cue to tonal contrast: A case study of Mandarin tones 2 and 3. *The Journal of the Acoustical Society of America* 93, 2241-2243.

Steriade, D. (2012). Intervals vs. syllables as units of linguistic rhythm. Handout, EALING, Paris.

Weismer, Gary, and Berry, Jeff (2003). Effects of speaking rate on second formant trajectories of selected vocalic nuclei. *Journal of the Acoustical Society of America* 113, 3362-3378.

Welby, P. (2006). French intonational structure: Evidence from tonal alignment. *Journal of Phonetics* 34, 343-371.

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. & 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, E.Q. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 33, 319–337.