Available online at www.sciencedirect.com
ScienceDirect
Journal of Phonetics 36 (2008) 724–746
Journal of Phonetics
www.elsevier.com/locate/phonetics

Emphasis and tonal implementation in Standard Chinese

Yiya Chen*, Carlos Gussenhoven

Radboud University Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Received 8 November 2006; received in revised form 17 June 2008; accepted 24 June 2008

Abstract

Despite the greatly improved understanding of tonal articulation in Standard Chinese, no consensus has been reached on the most appropriate model of tonal implementation [Xu, Y., & Wang, Q. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, *33*, 319–337; Kochanski, G., & Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, *39*(3/4), 311–352]. To shed new light on the issue, all four lexical tones, embedded in sentences with different preceding and following tonal contexts, were elicited under corrective focus, with two degrees of emphasis (Emphasis and MoreEmphasis), in addition to a NoEmphasis base-line condition, so as to bring systematic variation in duration and F_0 to bear on the issue of tonal realization in different pragmatic contexts.

Results showed comparable increases in syllable duration from the NoEmphasis condition to the Emphasis condition and from the latter to the MoreEmphasis condition. F_0 range expansion, however, was non-gradual: while there was a substantial increase in the F_0 range from the NoEmphasis to the Emphasis condition, the expansion from the Emphasis to the MoreEmphasis condition was marginal. Analyses of the F_0 patterns revealed that under emphasis, lexical tones were realized with magnified F_0 contours which were adapted to both the neighbouring tones and the durational increase of the tone-bearing syllables, and therefore maximally distinguishable from each other. Implications of these findings on models of tone and focus realization are discussed.

1. Introduction

Recent research has greatly increased our knowledge of the realization of the four lexical tones and the neutral tone of Standard Chinese (SC). In particular, the influence of preceding and following tones on the fundamental frequency has been documented in considerable detail (Chen & Xu, 2006; Shen, 1990; Shih, 1988; Xu, 1997, among others), while the effect of focus has also been subject to a number of investigations (Chen, 2003; Chen & Braun, 2006; Jin, 1996; Shih, 1988; Xu, 1999; Yuan, 2004). Despite this improved understanding, no consensus has been reached on how the tones are to be implemented in pragmatically different contexts. On the one hand, different views on the implementation of the tones are due to different conceptions of how they are represented in the phonology and thus to different conceptions of the relation between the phonology and the phonetics. On the other hand, these different views are also due to the different

*Corresponding author. Current address: Phonetics Laboratory, Cleveringaplaats 1 P.O. Box 9515, 2300 RA, Leiden, The Netherlands. Tel.: +31 71 527 1688; fax: +31 71 527 7569.
E-mail addresses: yiya.chen@let.ru.nl, yiya.chen@let.leidenuniv.nl (Y. Chen), C.Gussenhoven@let.ru.nl (C. Gussenhoven).

0095-4470/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2008.06.003

conceptions of how linguistic specifications (i.e. lexical tones in our case) are packaged or modified to convey pragmatic meanings. We believe that new data, in particular carefully elicited phonetic data with systematic variations, may well provide the supreme arbiter, much in the way that squishing and stretching complex objects may reveal the dimensions and elements of their structure.

The goal of this paper is two-fold. One is to shed further light on models of the prosodic endcoding of focus in tonal languages. The second is to test recent models for the phonetic implementation of lexical tones against new data collected under different emphasis conditions. To this end, we will compare the realization of lexical tones both outside the focus and under corrective focus. This difference is illustrated for the English word *tree* in (1) and (2), respectively.

(1) A: How did John say the word 'tree'?
B: John said the word '*tree*' slowly.

(2) A: John said the word 'flower' slowly.
B: No, John said the word '*tree*' slowly.

From a general communicative point of view, *tree* in the response in (1) represents the lowest degree of significance or emphasis, since it repeats information in the question by A, and may thus serve as a baseline pronunciation. By contrast, *tree* in the B-response in (2) represents a single corrected element in the preceding statement by A, and introduces a high degree of emphasis. The difference between the two pronunciations of *tree* in (1) and (2) is communicatively discrete, in the sense that a word is either a corrected element or is not. A gradient increase in communicative significance can be achieved by complicating the exchange in (2) as in (3), where speaker A at first misunderstands the reply by B.

(3) A: Did John say the word 'flower' slowly?
B: No, John said the word '*tree*' slowly.
A: Did you say he said the word '*bee*' slowly?
B: No, no, John said the word '*tree*' slowly!

The difference between the pronunciations of *tree* in the first and second statements by B in (3) is assumed to be gradient, and represents an increase in the care with which the word is pronounced. The three conditions may be referred to as NoEmphasis (1), Emphasis (2) and MoreEmphasis (3). Based on what is known about the realization of lexical tones in Standard Chinese, conditions like these may be expected to lead to considerable variation in acoustic parameters such as fundamental frequency and duration. Data collected under these three conditions are likely to allow us evaluate the predictions made by the models that have been proposed for phonetic implementation of lexical tones as well as models on the prosodic manifestations of focus in tonal languages.

1.1. Focus, degrees of emphasis and prosodic manifestations

For Germanic languages, a distinction is generally recognized between the effect of focus on the phonological representation and the effect of different degrees of emphasis on the phonetic implementation of the phonological tones. Presence of pitch accents on focused constituents and absence of pitch accents on postfocal constituents is commonly considered a structural device in the expression of focus (Gussenhoven, 1983; Ladd, 1980, 1996; Selkirk, 1984, 1995; cf. Xu & Xu, 2005). A gradual expansion of F_0 range for higher degrees of emphasis, however, is usually considered to be due to phonetic implementation (Rietveld & Gussenhoven, 1985; Ladd & Morton, 1997; Liberman & Pierrehumbert, 1984). While F_0 peak raising has been consistently observed, there is less clarity about the effect of emphasis on the scaling of F_0 valleys, which may be lowered (Rietveld & Gussenhoven, 1985; Liberman & Pierrehumbert, 1984; Gussenhoven & Rietveld, 2000, as measured with perceived surprise) or raised (Arvaniti & Garding, 2007). Thus, together with peak raising, the former results in F_0 span expansion, while the latter results in overall F_0 level raising, in the terms used in Ladd (1996). Interestingly, Arvaniti and Garding also report that the H in L* + H tone was aligned increasingly later with more emphasis, lending further support to the proposal that peak delay can be an effective substitute for peak raising to convey emphasis (Gussenhoven, 2004).

In addition to F_0, emphasis also induces durational increase (e.g. Sluijter, 1995; Turk & Sawusch, 1997). Arvaniti and Garding (2007) found that although duration increases consistently with higher degrees of emphasis, the magnitude of the increase is barely large enough to make them perceptible, given a JND of 10–40 ms, as reported in Lehiste (1970). This led them to conclude that at least in the task of their study, 'speakers relied more on F_0 and less on duration to indicate emphasis' (Lehiste, 1970, p. 21).

Chinese employs F_0 variation to indicate lexical tonal contrasts, and thus differs greatly from Germanic languages in this respect. Given that the functional load of F_0 in SC lies, to a very large extent, in word identification, it is plausible, following Berstein (1979), that while speakers of non-tonal languages such as English rely more on F_0 and less on duration to signal degrees of emphasis, SC is more restricted in the manipulation of F_0, relying more on duration. Salient duration increases to express degrees of emphasis, however, do not necessarily exclude the possibility that F_0 movements are additionally employed for this purpose. F_0 span expansion predicts that as the duration of the tone-bearing syllable increases, there should be increasingly higher F_0 maxima and lower F_0 minima. Also, F_0 level raising, causing both F_0 maxima and minima to be higher, will remain possible without necessarily destroying the characteristic contours of the lexical tones. By contrast, peak alignment (particularly in the Falling or Rising tones) may be adjusted only to the extent that the canonical lexical tonal shapes are not compromised, because tonal contours are lexically contrastive in SC. The question that we address, then, is how the SC lexical tones are implemented phonetically when uttered with greater emphasis, and specifically, if the duration of the tone-bearing syllable increases greatly as we have expected, how exactly the F_0 contours are adjusted.

1.2. Existing tone and intonation models

There have been quite a few models proposed in the literature on the interaction of tone and intonation in SC (Chao, 1968; Gårding, Zhang, & Svantesson, 1983; Kochanski & Shih, 2003; Xu, 2005; Xu & Wang, 2001; Wu, 1982; Yuan, 2004). With regard to the effect of focus (or emphasis) in SC, pitch range manipulation has been an important aspect of all models. Although Chao (1968) is the first to state that pitch range may be stretched or shrunken, similar to durational adjustment of syllables as longer or shorter, Gårding et al. (1983), to our knowledge, are the first to formalize pitch range manipulation in SC as the effect of *Range Grid*. The grid can be expanded (when under focus) or compressed (when out of focus) and therefore affects the pitch range of the lexical tonal realization in communicative contexts. Shih (1988), Jin (1996) and Xu (1999) have all adopted the notion of range expansion and suppression to account for the effect of focus or emphasis on tonal realization. In the following, we will focus on the two more recent models of tonal realization which make somewhat different predictions with regard to the interaction between F_0 contour and the increase in syllable duration.

The first is an *articulatorily* oriented model, proposed in Xu and Wang (2001) and later developed as the parallel encoding and target approximation model (PENTA) in Xu (2005). This model tries to simulate the articulatory process underlying the generation of surface F_0 contours. In this model, a lexical tone under focus is realized with two types of pitch targets. One type is local pitch target, which is determined by the lexical tone, and the other type is non-local pitch target which specifies 'the pitch range over which local pitch targets are implemented' (Xu & Wang, 2001, p. 334). The four lexical tones constitute the local pitch targets in SC. The High and Low tones have static pitch targets, while the Rising and Falling tones have dynamic targets. A local pitch target is implemented 'in synchronization with the tone-bearing syllable, i.e. starting at its onset and ending at its offset. Throughout the duration of the host, the approximation of the pitch target is continuous and asymptotic' (Xu & Wang, 2001, p. 322). The transition of one lexical tone to the next therefore starts at the onset of the syllable. This also predicts greater carry-over effects of a lexical tone on its following tone than anticipatory effects of the following tone on the preceding one.

Similar to local tonal targets, non-local 'pitch range targets are implemented asymptotically within its assigned temporal domain … If the domain is large, then the target can be reached well before the end of the domain … If the domain is small, then dynamic patterns similar to that of a local pitch target may result.' (Xu & Wang, 2001, p. 334). In a way, the non-local targets are similar to the *Range Grid*, and the PENTA model is along the line of the superposition accounts of intonation (Fujisaki, 1988; Gårding, 1979; Gårding et al., 1983; Grønnum, 1995, among others). Although Xu and Wang (2001) state explicitly that the domain of a non-local target is not determined by the pitch range target itself, they do not specify how the temporal domain of the non-local pitch targets is determined.

The second model, a *physiologically* motivated one, is the Soft TEMplate Mark-up Language (Stem-ML), proposed in Kochanski and Shih (2003). This model assumes that 'the prosodic trajectory is continuous and smooth over short time scales' (Kochanski & Shih, 2003, p. 315) given that prosody is controlled by muscle actions and muscles cannot discontinuously change positions. It employs the concept of tags which specify, among other things, a range value that sets the speaker's pitch range (i.e. the *range* attribute of the *set* tag), and a strength value that correlates with the linguistic strength of the tonal template (i.e. the *strength* attribute of the *stress* tag). Tonal strength also determines how one tone should interact with its neighbours; strong tones retain their shapes while weak ones accommodate the stronger ones. Consequently, strong tones keep their shapes more precisely than weak ones.

As discussed in Kochanski and Shih (2003), linguistic concepts like 'emphasis' can be expressed with the *range* attribute in a matrix. It is not explicitly stated whether in addition to pitch range, the strength with which the lexical tone is realized is also greater when under focus than in post-focus condition. But their general assumption that range is proportional to strength to the power of *ascale* (Kochanski, personal communication) suggests that the strength of an emphasized tone is expected to be higher than an un-emphasized one.

A third possibility can be deduced from a model that attempts to understand the possible universal and language specific aspects of tonal realization across both tonal and intonational languages. It assumes that all SC lexical tones are composed of high and low tones and realized as sequences of static tonal targets. In this model, the four tones of SC can be represented as H (for High tone), LH (for Rising tone), L (for Low tone) and HL (for Falling tone) (as in Duanmu, 2000), or phonetically more detailed such as (H)HH (for High tone), (L)LH (for Rising tone), (L)LL (for Low tone); (H)HL (for Falling tone) (as in Shih, 1988). Interpolations between the targets result in the rising and falling contours. With emphasis, in general, 'high targets become much higher, while low targets remain at the same level or are slightly lower' (Shih, 1988, p. 93). Hereafter, we will refer to this third possibility as the StaticTarget model.

In short, all three models predict F_0 range modification of tones for emphasis. PENTA predicts that such modification under focus should be restricted to the F_0 range designated specifically for the pragmatic function of focus. For more emphasis, PENTA predicts no further increase in range expansion given that the Emphasis and MoreEmphasis conditions are both to be categorized as focused condition. By contrast, Stem-ML and StaticTarget predict no specific pitch range limitation and therefore under different degrees of emphasis, they only expect possible ceiling effect of pitch range expansion due to the physiological limitation in the F_0 range of individual speakers. Given that it is much easier to raise F_0 maximum than to lower F_0 minimum, we expect maxF_0 to be a better indicator of the ceiling effect of pitch range expansion.

There is also a more general difference in perspective between PENTA and StaticTarget on the one hand and Stem-ML on the other. Focus realization is manifested in PENTA and StaticTarget predominantly via pitch range manipulation. In Stem-ML, by contrast, greater emphasis in principle can be modelled as the increase of effort in retaining the tonal shape. That is, there is an increase in hyperarticulation in the sense of Lindblom's (1990) H&H model in shifting from lesser to greater emphasis. This general difference should also be reflected in the way lexical tones are realized. Specifically, Stem-ML predicts that lexical tones under emphasis should be realized not only with expanded F_0 range but also with F_0 contours that are maximally distinct from each other, i.e. with hyperarticulated tonal targets.

With regard to lexical tonal realization, while the High and Low tones are mainly reflected via F_0 maximum and minimum respectively, good indicators for the realization of Rising or Falling tones include not only the alignment of F_0 turning point (often the max- and min-F_0) with segmental landmarks but also how the rising/falling movements are realized (e.g., the distance between F_0 minimum and maximum and consequently, the slope of F_0 rising/falling). PENTA assumes that there is a continuous approximation towards the tonal target throughout the tone-bearing syllable. It therefore predicts that if the duration of the tone-bearing syllables increases as a function of the degree of emphasis, tonal targets should be aligned consistently away from the syllable onset but close to the syllable offset. Stem-ML does not predict a specific relation between tonal targets and the tone-bearing syllable boundaries. Instead, as the overall magnitude of the tonal gesture increases with emphasis, it predicts that tonal targets may move away from the syllable boundaries if the duration of the tone-bearing syllable increases at a greater rate. At the same time, Stem-ML also dictates that tonal gestures are hyperarticulated, which means that the distance between the F_0-maximum and minimum should be enlarged while maintaining distinctive F_0 rising/falling contours. StaticTarget predicts two possibilities depending on whether lexical tonal targets are aligned with segmental landmarks in terms of absolute or relative distance: either the absolute distance of the tonal targets to the syllable edges may remain the same; or the proportional distance of the tonal targets to the syllable edges may remain the same.

The effect of the lexical tone on the pronunciation of preceding and following tones is also predicted to be different in the different models, given their different views on tonal coarticulation. We will take the offset F_0 of the preceding syllable (hereafter F_0-p) and the onset F_0 of the following syllable (hereafter F_0-f) as indicators of the effect of a target tone on its neighbouring tones. PENTA predicts more effect of the target tone on F_0-f than on F_0-p; Stem-ML predicts symmetrical effects of the target tone on F_0-f and F_0-p; but StaticTarget does not predict any direct effect.

To summarize, the goal of the paper is two-fold. One is to investigate how degrees of emphasis affect the duration and F_0 of the lexical tones of SC. Specifically, we were interested in the interaction of duration and F_0, so as to understand better how lexical tones are phonetically implemented. Our reference frame here is formed by the predictions that the three models of tonal implementation make about how emphasis-induced durational increase affects the F_0 contours. In so doing, we hope to obtain new data which will, in turn, shed new light on the strength and weakness of the models.

2. Method

2.1. Test materials

The target syllable in our materials is indicated as *Y* in the template sentence in (4), while the preceding and following syllables are identified as *X* and *Z*, respectively. For *Y*, all four lexical tones (i.e. High, Low, Rising and Falling) were included. As shown in Table 1, its syllable structure includes both simple CV (i.e. *ma*) and complex CGVG (i.e. *miao*). The preceding syllable *X* varies between *shuō* (to 'say') with a High tone (h) and *xiĕ* ('to write') with a Low tone (l). The following syllable *Z* varies between *nán* ('difficult') with a Rising tone (lh) and *màn* ('slow') with a Falling tone (hl). As a result, target syllable *Y* may be preceded by tones that end high or low and followed by tones that start high or low. The lexical items for *X*, *Y* and *Z* were chosen so as to obtain sentences that were readily interpretable, with the desired tonal sequences on syllables in the *X*, *Y* and *Z* positions that were easy to segment, and whose segmental structures were identical, or if this was not possible, comparable. Sixteen stimulus sentences were included.

(4) zhōu bīn shuō X Y Z hĕn duō.
zhōu bīn said X Y Z very more
'zhōu bīn said it is much more *Z* (difficult/slow) to *X* (write/say) *Y* (target syllable with four different tones).'

2.2. Discourse context

The stimulus sentences were elicited in three Discourse contexts: NoEmphasis, Emphasis and MoreEmphasis. The NoEmphasis condition served as a baseline for the Emphasis and MoreEmphasis

Table 1
Stimulus set

X	*Y*	*Z*
shūo (h) xiĕ (l)	ma/miao (4 lexical tones)	nán (lh) màn (hl)

conditions. In the Emphasis condition, subjects were first given a sentence in Chinese characters on a computer screen, labelled ‘correct information’. An example in pinyin is given in (5). Subsequently, they were given a new sentence in which one element was different, this time with the label ‘incorrect information’, as illustrated in (6), again in pinyin. On the same screen, they were given the instruction to provide a correction, as in (7). A typical answer from the speakers, with emphasis on *miao* (bold and underlined), is illustrated in (8).

(5) *Correct information:*
zhōu bīn shuō shuō miāo nán hěn duō.
‘Zhoubin said that it is more difficult to say *miao*.’
(6) *Incorrect information:*
zhōu bīn shuō shuō dǎ nán hěn duō.
‘Zhoubin said that it is more difficult to say *da*.’
(7) *Context for Emphasis:*
Suppose you gave the correct information in sentence (4), and the experimenter thought you said sentence (5), how would you correct the experimenter?
(8) *Response with emphasis:*
zhōu bīn shuō shuō **miāo** nán hěn duō.

For the MoreEmphasis condition, the experimenter pretended that she had not heard the subject clearly, and asked for a repetition. This led the subject to say (8) again with greater emphasis on the syllable *miao*, as indicated with double underline in (9).

(9) *Response in MoreEmphasis condition:*
zhōu bīn shuō shuō **miāo** nán hěn duō.

The baseline condition, NoEmphasis, was similarly elicited with the help of a question about the sentence spoken by the subject, but this time the question sought to obtain the information expressed in the adverb, as illustrated in (10). A typical answer will not have any emphasis on the target syllable, and instead, will be on the last syllable. This is illustrated in (11). There were two subjects who sometimes produced *nán duō le* instead of *nán hěn duō*. Since this alternation should not (and indeed did not) affect the production of the target syllable significantly, they were not corrected.

(10) *Baseline (NoEmphasis) condition*:
zhōu bīn shuō shuō miāo zěnme yàng?
What did Zhoubin say about saying the word *miao*?
(11) *Response with no emphasis*:
zhōu bīn shuō shuō miāo nán hěn duō.

2.3. Subjects and recording

Three male and two female speakers of SC participated in the experiment. Four were born and raised in Beijing. One was born elsewhere, but grew up in Beijing and speaks SC without any detectable accent, as judged by the first author and two other native speakers. All speakers were recorded in a sound-treated booth, three at Stony Brook University, with a Sony Digital Mega Bass MZ-R55 mini recorder, and two at Radboud University Nijmegen, with a DAT-recorder. All other conditions were kept the same. Data from the three Stony-Brook subjects have been reported in Chen (2003) and Chen (2006). The whole data set of all five subjects was completely reprocessed and reanalyzed.

The test materials were presented to each of the five subjects three times. The 32 sentences were automatically randomized on each presentation. Recordings were made first at a sampling rate of 44,100 Hz, and then down-sampled to 16,000 Hz. All speakers were aware that it was a study of prosody in SC, but were naïve as to the specific purpose. During the recording, speakers were asked to repeat the sentences whenever the experimenter failed to perceive the intended pragmatic meaning.

2.4. Acoustic analysis

The start and end of each of the syllables designated *X*, *Y* and *Z* were manually labelled in Praat (Boersma & Weenink, 2005), and durations were obtained. In addition, several F_0 measurements were made for statistical analyses. For the preceding syllable *X*, we took the F_0 offset (F_0-p) and for the following syllable *Z*, we took the F_0 onset (F_0-f). For the target syllable *Y*, we measured the F_0 maximum (maxF_0) and minimum (minF_0) values and their locations in all four lexical tones. As illustrated in Fig. 1A and B, for the Rising tone, minF_0 was searched first and then maxF_0; while for the Falling tone, maxF_0 was marked before minF_0. For brevity, we will from here on use H, LH, L and HL for the High, Rising, Low and Falling lexical tones of SC, without committing ourselves to a particular phonological representation.

In most cases, F_0 minima and maxima were easy to identify, and reflected the turning point of LH and HL. For example, when LH was preceded by H, the start of the rise corresponded roughly to minF_0. When LH was preceded by L, however, there was sometimes a low elbow in which the location of the minF_0 could be some distance away from the turning point, as illustrated in Fig. 2. We thus introduced another variable, the start of rise (R), which was indicative of how the rising contour was realized. Mathematically, the start of rise was defined as the time point of the intersection of two straight lines: one is the tangent at the maximum second derivative located after minF_0 (i.e. the maximum acceleration of the F_0 rising curve), and the other is the tangent at the lowest positive second derivative between the F_0 minimum and the maximum second derivative. In our data set, this measure corresponded the best with the F_0 rising turning point by eye-balling. The R in Fig. 2 illustrates its typical approximate location. While R was often later than minF_0 after L, its F_0 was nevertheless close to minF_0.

For HL, we always took maxF_0 as the start of fall as it was quite easy to identify and corresponded well with the start of falling F_0. In the case of L, minF_0 reflects only the lowest point of *the measurable part* of the F_0 contour, as many tokens exhibited varying degrees of creakiness or even glottalization. Impressionistically, creakiness varied with speaker as well as with the experimental emphasis condition. Among the five speakers, four pronounced L with more creakiness in the MoreEmphasis condition than in the Emphasis condition, while there was much less creakiness in the NoEmphasis condition. One female speaker showed creakiness in all three conditions.

The f_0 values in Hz were converted into semitones to reduce cross-speaker variation. Formula (1) relates frequency in semitones, *F*, to frequency in Hz, *f*:

$$F = 12\log_2\left(\frac{f}{100}\right). \tag{1}$$

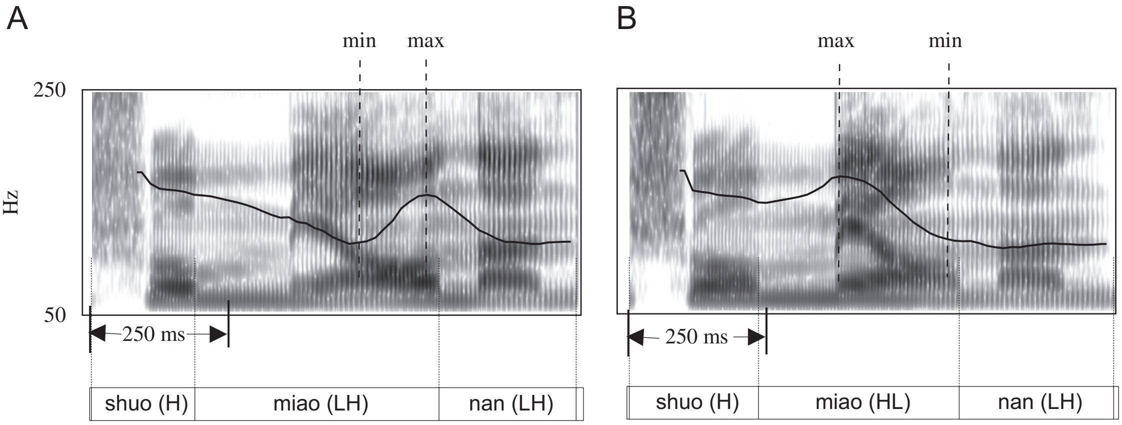

Fig. 1. Spectrograms with superimposed f_0-contours showing f_0 minima (min) and maxima (max) in 'shuo (H) miao (LH) nan (LH)' (left) and 'xie (L) miao (HL) nan (LH)' (right) produced by an adult male speaker. (The frequency range of the spectrograms is 0–5000 Hz.)

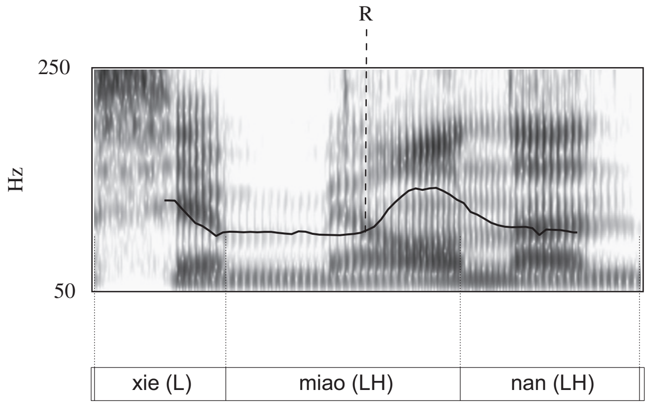

Fig. 2. Spectrograms with superimposed f_0-contour showing the start of f_0 rise (R) in 'xie (L) miao (LH) nan (LH)' produced by an adult male speaker. (The frequency range of the spectrogram is 0–5000 Hz.)

Syllable duration, S_{dur}, is the duration from the acoustic onset to the acoustic offset of the syllable. The range in semitones, F_{range} is given as:

$$F_{\mathrm{range}} = 12\log_2\left(\frac{f_{\max}}{f_{\min}}\right), \tag{2}$$

where, $f_{\max}$ and $f_{\min}$ are the maximum and minimum frequencies in Hz.

Tone$_{\mathrm{dur}}$, the absolute duration between $t_{\max}$, the time at which $f_{\max}$ occurs, and $t_{\min}$, the time of $f_{\min}$:

$$\mathrm{Tone}_{\mathrm{dur}} = |t_{\max} - t_{\min}|. \tag{3}$$

P_{dur}, the proportional time taken up by f_0-change in the tone relative to the total syllable duration:

$$P_{\mathrm{dur}} = \frac{\mathrm{Tone}_{\mathrm{dur}}}{S_{\mathrm{dur}}}. \tag{4}$$

Linear slope of the f_0 rise or fall:

$$F_{\mathrm{slope}} = \frac{F_{\mathrm{range}}}{\mathrm{Tone}_{\mathrm{dur}}}. \tag{5}$$

Lag$_{\mathrm{dur}}$, the duration from the syllable onset S_{onset} to the onset time of the f_0 rise or fall Tone$_{\mathrm{onset}}$:

$$\mathrm{Lag}_{\mathrm{dur}} = \mathrm{Tone}_{\mathrm{onset}} - S_{\mathrm{onset}}, \tag{6}$$

where Tone$_{\mathrm{onset}}$ and S_{onset} are the onset time of the rising/falling tone and syllable, respectively.

Lag$_{\mathrm{p}}$, the proportional time at which the f_0 fall or rise begins relative to the total syllable duration:

$$\mathrm{Lag}_{\mathrm{p}} = \frac{\mathrm{Lag}_{\mathrm{dur}}}{S_{\mathrm{dur}}}. \tag{7}$$

2.5. Statistical analysis

Our goal was to examine the effect of the discourse condition on a number of duration and F_0 measurements that characterize the realization of the lexical tones in ways that might be testable against the three models for the phonetic implementation of tones and focus realization in tonal languages. Unless otherwise noted, Repeated Measures Analyses by Subjects were conducted on these variables with the factors DISCOURSE (three levels: NoEmphasis, Emphasis and MoreEmphasis), TONE (four levels: High, Low, Rising and Falling), SYLLABLE STRUCTURE (two levels: complex and simple), PRECEDING TONE (two levels: High and Low) and FOLLOWING TONE (two levels: Rising and Falling). Our main interest was in the effect of DISCOURSE and its possible interaction with other factors. So, significant interactions of factors other than those with DISCOURSE may not be covered (especially when the magnitude of such interactions was ordinal and negligible).

3. Results

3.1. Graphical comparison of the F_0 contours

Fig. 3 displays mean F_0 contours of the four lexical tones (indicated by the first letter of the names in the legend), preceded by High (left column) or Low (right column), and followed by Rising or Falling (indicated by the second letter of the names in the legend), uttered in the NoEmphasis (N), Emphasis (E) and MoreEmphasis (M) conditions (indicated by the last letter of the names in the legend). The grey lines here indicate NoEmphasis, the black solid lines indicate the Emphasis condition, and the black dotted lines the MoreEmphasis condition.

These F_0 contours were obtained by taking 20 F_0 points (in Hz) at proportionally equal time intervals between the acoustic onset and offset of the syllable and averaging these across three repetitions of the same sentence for each emphasis condition separately. These values were then transformed into semitones and averaged across speakers. Since the F_0 contours for the two syllables (i.e. *miao* and *ma*) were essentially the same, we pooled across these contours, for each tone separately. The normalized syllable duration is the mean duration of the target syllable averaged across speakers, repetitions and syllable structures.

Three things are to be noted. First, the Emphasis condition indeed induced a wider F_0 range than the NoEmphasis condition with F_0 peaks being raised more than F_0 valleys being lowered. The difference in F_0 range between Emphasis and MoreEmphasis, however, was less pronounced, and varied across tones. While LH and HL showed differences between these two conditions, there were only marginal differences for H and L.

Second, each tone exhibited a distinctive pattern of F_0 movement characteristic of the tonal feature(s). H was produced with a raised F_0 peak. L was realized with a slightly lowered F_0 value (with creakiness or glottalization, not shown here, which may be taken as an enhancement of low). LH and HL exhibited higher F_0 peaks, slightly lowered F_0 valleys, and a clearly delayed start of the rise and the fall, which resulted in F_0 shapes for LH and HL that were distinctively different from those for H and L.

Third, there was a robust effect of the preceding tone on the F_0 contours of H, LH and HL, as can be seen by comparing the left and right columns in the same row. Specifically, the F_0 contours of these lexical tones started high when preceded by H and low when preceded by L. The effect of the preceding tone on L was more subtle, however, and L started high both after H and after L. This suggests that even when a Low tone was emphasized, Third-tone sandhi, which changes the first of a sequence of two Low tones into LH, applied here.

In Section 3.2, we will report the results of a number of quantitative analyses. The application of Third-tone sandhi, together with the fact that L tones were realized with considerable glottalization, led us to leave out L from statistical analyses of F_0 measurements.

3.2. Quantitative analyses

3.2.1. Syllable duration lengthening

The duration of the target syllable (S_{dur}) was significantly affected by DISCOURSE [$F(2, 8) = 125.9$, $p < .0001$]. Bonferroni post-hoc tests showed that all three levels of emphasis differed significantly from each other. As shown in Fig. 4, the target syllable in the Emphasis condition was on average 81 ms longer than that in the NoEmphasis condition, an increase of 34%. The MoreEmphasis condition caused a further increase of 87 ms, an increase of 27% relative to the Emphasis condition. By contrast, neither the preceding nor the following syllable exhibited comparable changes in duration, suggesting that the durational increase in the target syllable cannot be due to any adjustment of the speaking rate.

There was also a significant main effect of TARGET TONE [$F(3, 8) = 11.6$, $p < .01$] and FOLLOWING TONE [$F(1, 4) = 23.6$, $p < .01$] on the duration of the target syllable. Bonferroni post-hoc tests showed that the only significant difference was that the High tone was on average 9 ms shorter than the LH tone. Second, target syllables preceding LH were significantly longer (9 ms) than target syllables preceding HL.

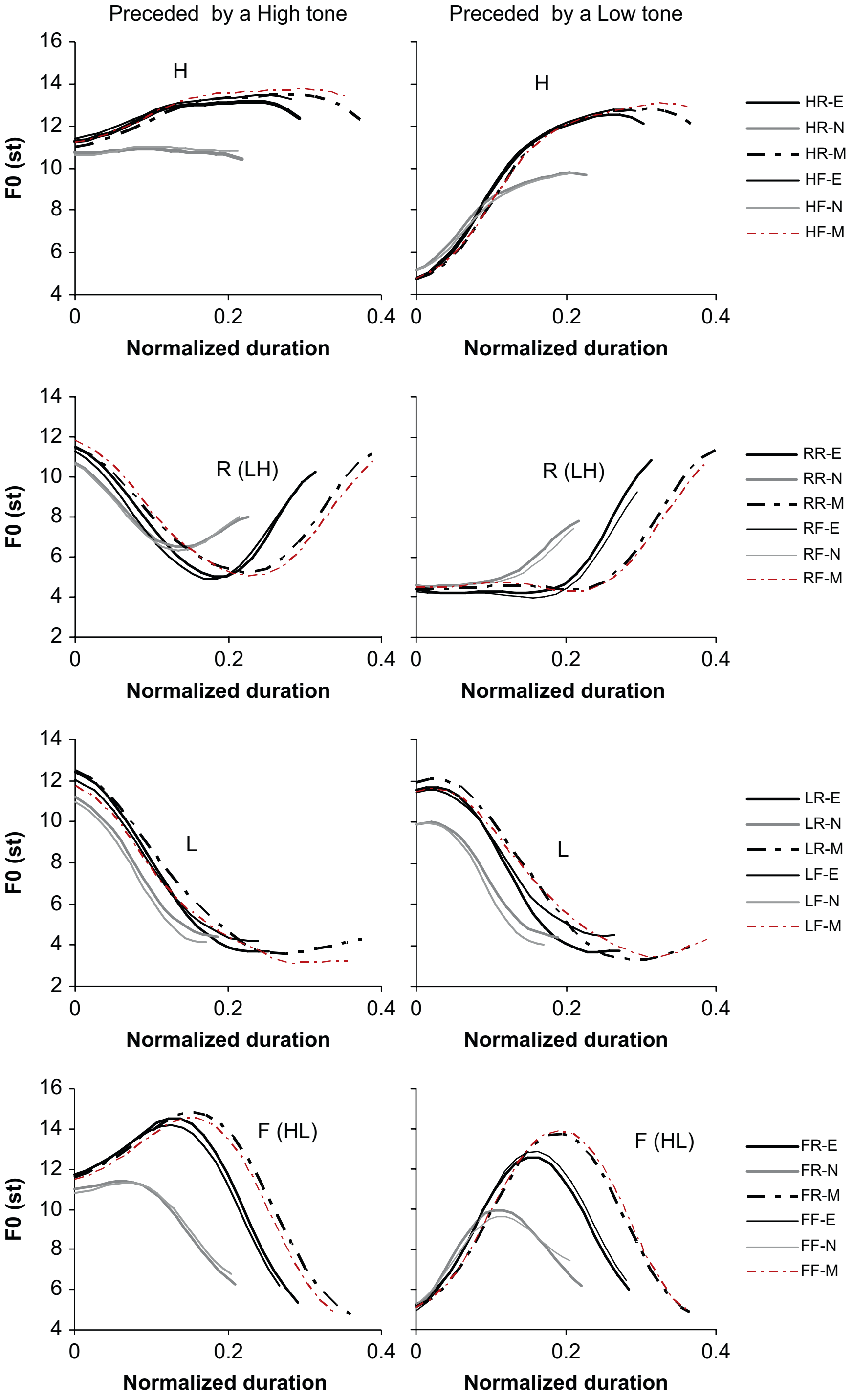

Fig. 3. F_0 contours that were averaged separately across speakers and repetitions for the various conditions shown after linear time normalization. The four lexical tones are High (H), Rising (R), Low (L) and Falling (F) followed by either a Rising (R) or a Falling (F) tone, uttered with different emphasis conditions (N = NoEmphasis, E = Emphasis and M = MoreEmphasis). The normalized syllable duration for each emphasis condition is the mean duration of the target syllable averaged across speakers, repetitions and syllable structures.

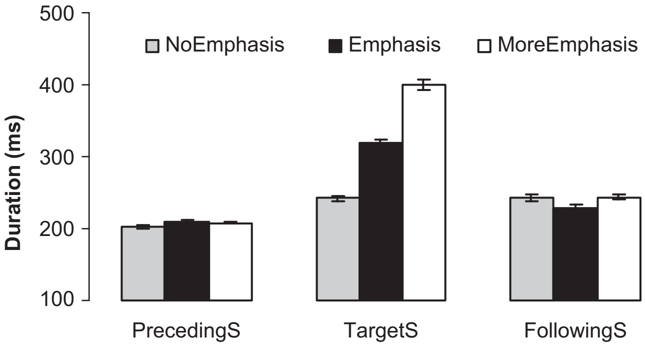

Fig. 4. Mean duration (and ±2 standard errors) of the target syllable, the preceding syllable and the following syllable under three emphasis conditions.

No significant interaction was found, suggesting that the effect of DISCOURSE was independent of other factors that affected syllable duration.

In short, corrective focus induced significant lengthening. Under corrective focus, the durational adjustment was a robust manifestation of two degrees of emphasis: the greater emphasis with which the corrected target syllable was produced, the longer was its duration.

3.2.2. F_0 range expansion

The F_0 range of the target syllable (F_{range}) was also significantly affected by DISCOURSE [$F(1.07, 4.28) = 34.5$, $p < .01$]. Bonferroni post-hoc tests showed that there was a significant difference between NoEmphasis and Emphasis, but no significant difference was found between the Emphasis and the MoreEmphasis conditions. On average, the F_0 range of the target syllable in the Emphasis condition was 3.1 st wider than that in the NoEmphasis condition, an increase of 102% (i.e. 125% of increase in absolute range in Hz). The F_0 range of the target syllable in the MoreEmphasis condition, however, only exhibited an increase of about 1.1 st, or 18% (i.e. 20% of increase in absolute range in Hz), relative to the Emphasis condition.

The magnitude of the F_0 range expansion was conditioned by all other factors: SYLLABLE STRUCTURE × TARGET TONE × DISCOURSE: [$F(4, 16) = 3.2$, $p < .05$]; PRECEDING TONE × TARGET TONE × DISCOURSE: [$F(4, 16) = 12.4$, $p < .0001$]; FOLLOWING TONE × TARGET TONE × DISCOURSE: [$F(4, 16) = 3.1$, $p < .05$]. All interactions were ordinal. The most important salient factors in determining the magnitude of F_0 range expansion were PRECEDING TONE and TARGET TONE, as illustrated in Fig. 5. For instance, a HL tone exhibited a wider F_0 range than a H tone[1], and a H tone had more scope to vary its pitch range after L than after H. In general, tones preceded by L had a wider F_0 range than those preceded by H (by 1.6 st). In addition, tones followed by LH showed a more expanded range than those followed by HL (by .35 st).

In sum, the important finding here is that the variation of F_0 as a function of DISCOURSE differed markedly from the variation in duration. While the Emphasis condition induced a substantial expansion of the F_0 range as well as an increase in syllable duration relative to the NoEmphasis condition, the MoreEmphasis condition only induced a much smaller and inconsistent increase in F_0 range relative to the Emphasis condition, even though the corresponding durational increase showed a statistically significant increase with comparable magnitude to the increase from the NoEmphasis to the Emphasis condition. Moreover, lexical tones exhibited different limits on the extent to which their F_0 range could be expanded, with HL expanding much more than LH and H.

[1] The F_0 range of a high tone was derived as the difference between the max F_0 realized as the high tone and the min F_0 within the high tone carrying syllable, which was usually the end F_0 of the preceding tone.

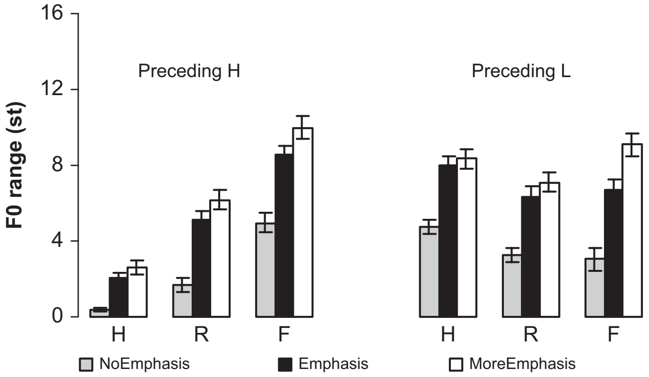

Fig. 5. Mean F_0 range (and ± 2 standard errors) of the three lexical tones (H: High; R: Rising; F: Falling) in three emphasis conditions, preceded by High or Low tone.

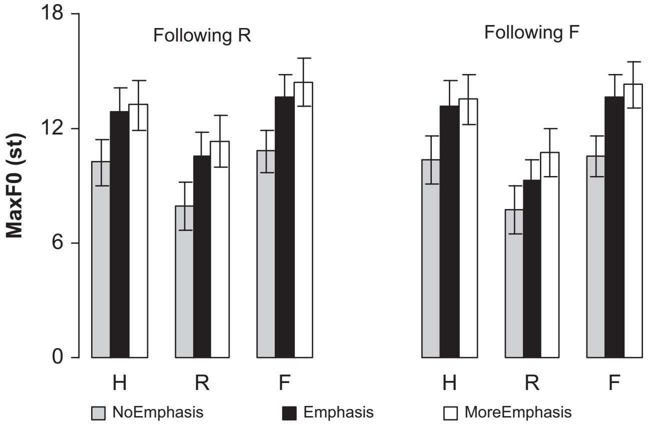

Fig. 6. Mean maxF_0 (and ± 2 standard errors) of the three lexical tones (H: High; R: Rising; F: Falling) in three emphasis conditions, followed by Rising or Following tone.

3.2.3. *Scaling of the F_0 maximum and minimum*

To investigate what gave rise to the F_0 range expansion, the scaling of the F_0 maximum and minimum of the target tones was also examined. We will report on the results of MaxF_0 first. There was a significant main effect of DISCOURSE [$F(2, 8) = 33.1$, $p < .0001$]. Bonferroni post-hoc tests showed that the F_0 range of the target syllable differed significantly only between the NoEmphasis and the two emphasis conditions. On average, maxF_0 in the Emphasis condition (12.2 st) was higher than in the NoEmphasis condition (9.6 st), an increase of 2.6 st (27%). In the MoreEmphasis condition, maxF_0 (12.9 st) tended to be higher than in the Emphasis condition, with an average increase of .6 st (5%).

DISCOURSE interacted significantly with TARGET TONE and FOLLOWING TONE in determining F_0 peak raising [$F(4, 16) = 3.96$, $p < .05$]. As illustrated in Fig. 6, TARGET TONE had a much greater effect on peak raising than FOLLOWING TONE. Bonferroni post-hoc tests showed that both HL and H tones had significantly higher F_0 peaks than the LH tone. Although no significant difference was found between HL and H tones, it is interesting to note that for most subjects, the F_0 peak over a HL tone tended to be higher than that over an H tone. Fig. 7 shows the difference in MaxF_0 between HL and H tones produced in the MoreEmphasis condition by each of the five subjects. In the upper panel, the following tone was LH and in the lower panel, it was HL.

As for MinF_0, only LH and HL tones were included, since the MinF_0 of a High tone (H) was determined by the end point of the preceding tone and thus did not form an intrinsic part of a High tone, as is also clear from

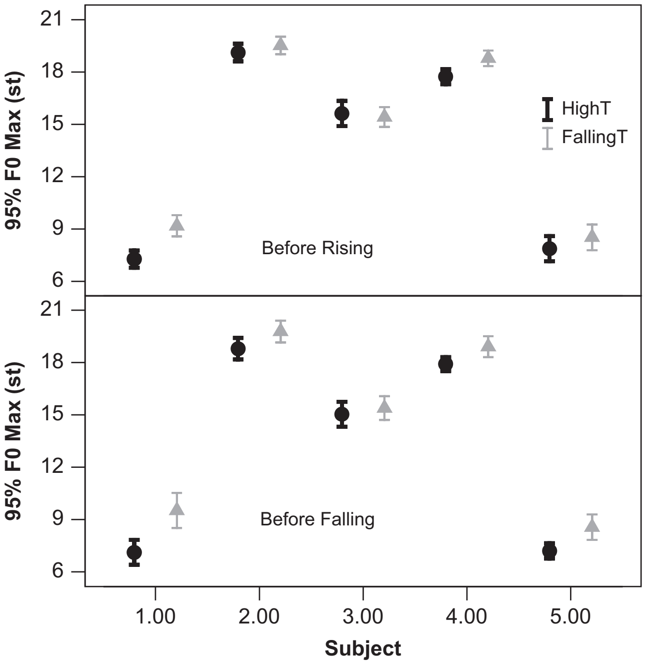

Fig. 7. Boxplots of the MaxF_0 of HL and H tones produced in the MoreEmphasis condition by each of the five subjects. In the upper panel, the following tone was LH and in the lower panel HL.

Fig. 3. DISCOURSE was a significant factor [$F(2, 8) = 8.2$, $p < .05$]. Bonferroni post-hoc comparisons, however, showed that there was no significant difference between any two of the three emphasis levels although the trend was that mean MinF_0 lowered with higher degree of emphasis (NoEmphasis: 6.0 st; Emphasis: 5.1 st; MoreEmphasis: 4.6 st).

To summarize, DISCOURSE significantly affected both MaxF_0 and MinF_0. Although there was a significant increase of MaxF_0 from NoEmphasis to Emphasis, no significant difference was found between Emphasis and MoreEmphasis. As for MinF_0, there was no significant difference between any two of the three emphasis conditions. The general tendency, however, was for MinF_0 to be lower and MaxF_0 higher as the degree of emphasis increased, giving rise to the expansion of the F_0 range discussed in Section 3.2.2. Furthermore, we also found a general pattern of tonal intrinsic MaxF_0. On average, HL showed the highest F_0 and LH the lowest. This also correlated with the tonal intrinsic F_0 range expansion that we have observed earlier.

3.2.4. F_0 rise and fall alignment

In order to understand better how F_0 contours of lexical tones as a whole were affected by the Emphasis and MoreEmphasis conditions, we further examined how the F_0 rising and falling movements were timed with reference to the tone-bearing syllable edges.

Fig. 8A and B shows that the start of the rise for a LH tone and of the fall for a HL tone correlated strongly with the increase of the syllable duration. This suggests that as the duration of the syllable increased, the start of F_0 rising or falling movements of both tones was delayed, which confirms the findings of Xu (1999) that the rising and falling movements are timed with reference to the offset of the syllable, rather than the onset.

The relative distance from the start of rise/fall to the onset of the tone-bearing syllable ($\mathrm{Lag}_{\mathrm{dur}}$) was significantly affected by DISCOURSE [$F(2, 8) = 14.8$, $p < .01$]. Bonferroni post-hoc comparisons showed that the start of F_0 movement was significantly later in the Emphasis (at 58.5% of the syllable duration) and MoreEmphasis conditions (56.4%) than in the NoEmphasis condition (51.4%), but there was no significant difference between the Emphasis and MoreEmphasis conditions.

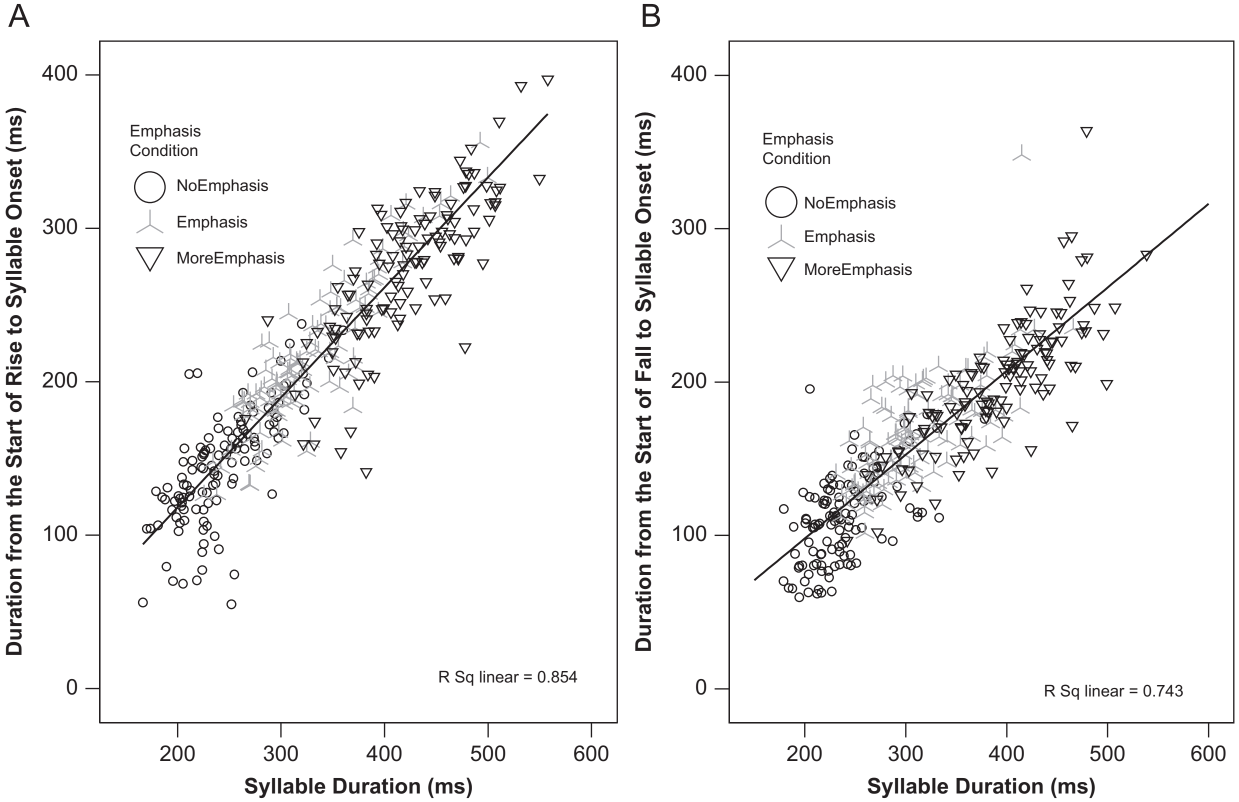

Fig. 8. Syllable duration as a function of the interval between the syllable onset to the onset of the rise (Fig. 10A) and to the onset of the fall (Fig. 10B) in three emphasis conditions with superimposed straight lines of best fit.

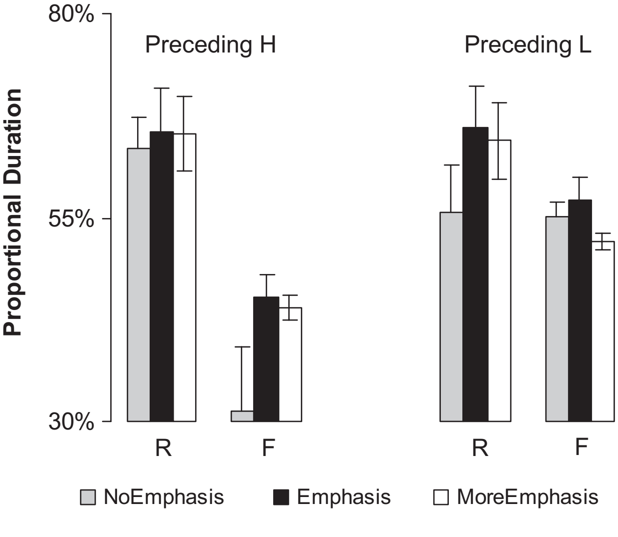

Fig. 9. Average proportional duration (%) between the onset of the rise/fall relative to the syllable onset as a function of total syllable duration.

Of particular note is the significant three-way interaction of DISCOURSE × PRECEDING TONE × TARGET TONE [$F(2, 8) = 16.2$, $p < .01$], as illustrated in Fig. 9. After H, the fall for HL began significantly earlier than after Low, because it took time for F_0 to rise from the low ending of L to the high beginning of

HL. However, the start of Rising was only significantly affected by the PRECEDING TONE in the NoEmphasis condition. In the Emphasis and MoreEmphasis conditions, PRECEDING TONE showed no effect, despite the fact that when preceded by a High tone, it must have taken time for F_0 to fall from the high ending of H to the low beginning of LH.

3.2.5. Rise/fall time and slope

Two other aspects were examined to shed further light on how lexical tonal contours were realized in different discourse contexts: (1) the distance between the maximum and minimum F_0 in both absolute duration (Tone$_{\mathrm{dur}}$) and proportional duration (P_{dur}); and (2) the slope of the F_0 rise and fall (F_{slope}).

Here, we selected a subset of data which showed clear low and high turning points. Specifically, for LH, we included only those token that were preceded by H and followed by LH; for HL, we included only those tokens that were preceded by L and followed by HL. Both absolute and relative distances were calculated (the latter being expressed as a proportion of the syllable duration) and subjected to Repeated Measures Analyses by Subjects with DISCOURSE (three levels: NoEmphasis, Emphasis and MoreEmphasis), TARGET TONE (two levels: LH and HL), and SYLLABLE STRUCTURE (two levels: complex and simple).

DISCOURSE had a significant effect on the absolute distance between the F_0 minimum and maximum [$F(2, 8) = 34.8, p < .0001$]. Bonferroni post-hoc comparisons showed that all three degrees of emphasis differed significantly from each other. As the degree of emphasis increased, the distance between MaxF_0 and MinF_0 increased, as shown in Fig. 10.

The *relative* distance of the movements also showed a significant main effect of DISCOURSE [$F(2, 8) = 10.6, p < .01$]. Bonferroni post-hoc comparisons, however, showed no significant difference between any two of the three emphasis conditions. In other words, as the degree of emphasis increased, there was only a tendency for the proportional duration of the fall and rise to increase from the NoEmphasis (34%) to the Emphasis (38%) and MoreEmphasis conditions (41%). This suggests that the distance between the F_0 maxima and minima increased along with the duration of the tone-bearing syllable as a whole, more or less proportionally with the increase in the syllable duration.

The *slope* of the F_0 rise and fall was significantly affected by DISCOURSE [$F(2, 8) = 23.2, p < .0001$] and TARGET TONE [$F(1, 4) = 9.7, p < .05$], the two of which also showed a significant interaction [$F(2, 8) = 9.5, p < .01$]. Bonferroni post-hoc comparisons showed a significant difference between NoEmphasis and the two emphasis conditions, but no significant difference was found between the Emphasis and MoreEmphasis

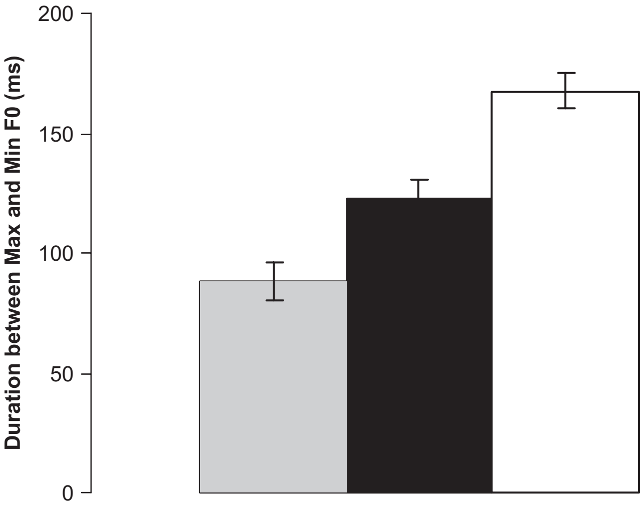

Fig. 10. Average duration between Max and MinF_0 of the Rising tone (preceded by a High tone and followed by a Rising tone) and Falling tone (preceded by a Low tone and followed by a Falling tone) (and ±2 standard errors).

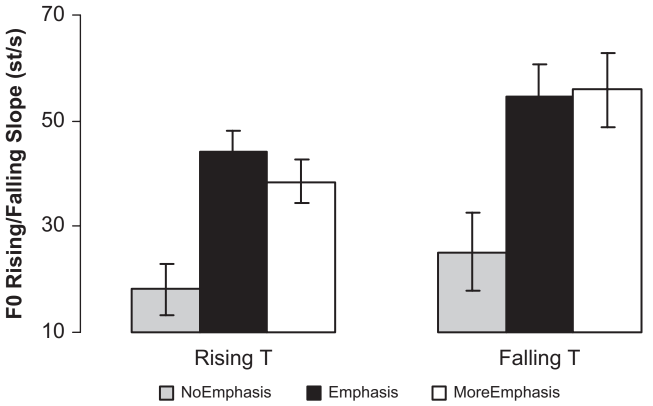

Fig. 11. Average linear slopes of the Rising and Falling tones in three different emphasis conditions.

conditions, as also illustrated in Fig. 11. Rising on average had a less steep F_0 slope than falling, which is consistent with the finding of Xu and Sun (2002).

3.2.6. The offset F_0 of the preceding tone and onset F_0 of the following tone

To understand the patterns of tonal transition between a target tone (*Y*) and its preceding tone (*X*) as well as its following tone (*Z*) under different degrees of emphasis, we further examined the F_0 offset of the preceding tone (F_0-p) and the onset of the following tone (F_0-f). As mentioned earlier, the Third-tone sandhi rendered the effects of the two different preceding tones (i.e. Low and High) comparable before the target Low tone, as the preceding Low tone was realized with a rising f_0 contour which ended with high f_0 like a High tone. Therefore, we excluded the Low tone from the statistical analyses but included it in the graphs.

For the preceding offset F_0-p, there was a significant effect of PRECEDING TONE [$F(1, 4) = 170$, $p < .0001$], TARGET TONE [$F(3, 12) = 97$, $p < .0001$], as well a significant interaction of these two factors [$F(3, 12) = 38$, $p < .0001$]. As shown in Fig. 12A, the offset F_0 value was high when the preceding tone was H and low when the preceding tone was L (except for the case of two Low tones). The effects of TARGET TONE and DISCOURSE were negligible. This suggests a rather limited anticipatory effect. The data also suggest an anticipatory dissimilation effect in that when the Target tone was Low or LH, the offset F_0-p tended to be higher.

For the following onset F_0-f, there were main effects of the TARGET TONE [$F(2, 8) = 98.2$, $p < .0001$], FOLLOWING TONE [$F(1, 4) = 8.0$, $p < .05$], DISCOURSE [$F(2, 8) = 8.8$, $p < .05$], as well as the interaction of TARGET TONE × DISCOURSE × FOLLOWING TONE [$F(4, 16) = 3.5$, $p < .05$]. As shown in Fig. 12B, after the High and LH tone, the onset F_0-f was raised greatly under the two emphasized conditions. As a contrast, after the Low tone, the onset F_0-f remained the same; after the HL tone, the onset F_0-f was slightly lowered. The difference between Emphasis and MoreEmphasis, however, was negligible.

The effects of TARGET TONE and DISCOURSE on the F_0 contours of the following tone extended beyond the onset F_0-f into the whole syllable, as shown in the schematic contours in Fig. 13. Here, F_0 values of five equi-distant points (P1–P5) over the following syllable *Z* (i.e. *nán* 'difficult' with a LH tone (A) and *màn* 'slow' with a HL tone (B)) were plotted. In each column, the LH (upper row) and HL (lower row) tones followed one of the four Target tones (i.e. H, LH, L and HL). Throughout the syllable, the LH and HL tones were both realized with F_0 contours which showed clear rising or falling shapes when the target tone was in the NoEmphasis condition (solid grey lines). When the target tone was in the Emphasis (solid black lines) and MoreEmphasis (dotted black lines) conditions, the LH tone was realized with a clear rising F_0 contour only when following the target Low tone; the HL tone was realized with a falling F_0 contour when following all four target tones but the magnitude of its fall varied depending on the height of the target tone's F_0 end point: it was greater when the target tone ended with a high F_0 (i.e. in the H and LH target tones).

In short, the investigation of the tonal transitions to and from the target lexical tones revealed significant and robust effects of TARGET TONE and DISCOURSE on the onset F_0-f of the following tone, but not on

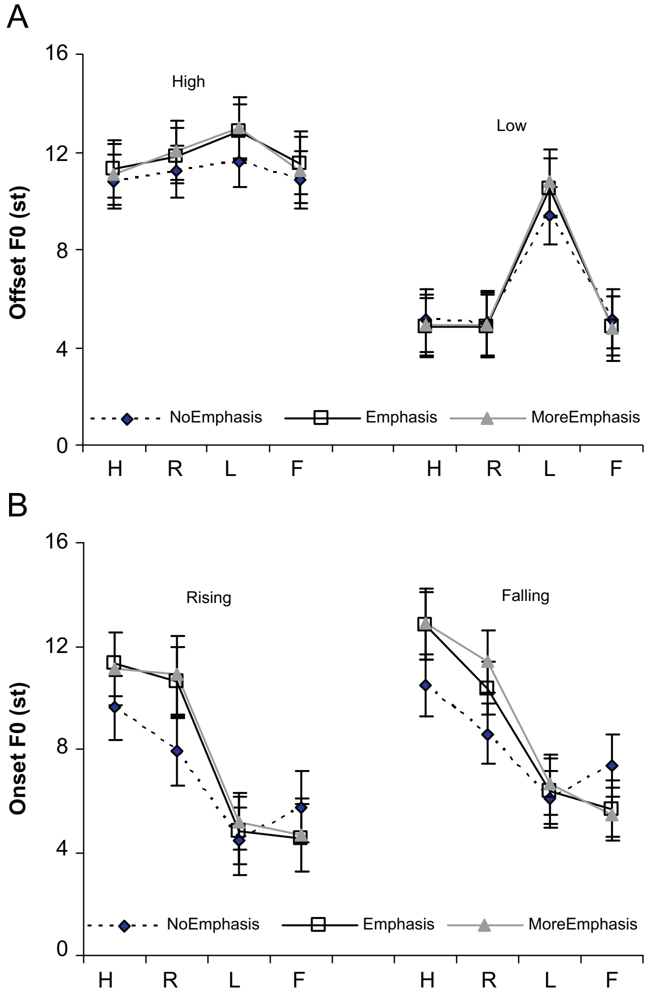

Fig. 12. Average F_0 (and ±2 standard errors) of the end point (Offset F_0) of the Preceding tone (A) and onset F_0 (Onset F_0) of the Following tone (B) as a function of different Target tones, Following tones, and emphasis conditions.

the offset F_0-p of the preceding tone. Furthermore, the effect of DISCOURSE and the co-articulation of the target and following tones were specific to the tonal contexts and lasted throughout the following syllable.

4. Discussion and conclusion

4.1. Duration and F_0 adjustment as function of degrees of emphasis

There was a robust and gradual increase in syllable duration from the NoEmphasis via the Emphasis to the MoreEmphasis condition. The F_0 range expansion, however, was non-gradual, as illustrated by the contrast in Fig. 4 (on duration) and 5 (on F_0 range). Specifically, although there was a robust increase of F_0 range from the NoEmphasis to the Emphasis condition, further expansion from the Emphasis to the MoreEmphasis condition was limited and showed no statistical significance. This makes it clear that in Standard Chinese, corrective focus does indeed induce a significant durational increase and F_0 range expansion, lending further support to the existing literature on focus realization (Chen, 2003; Jin, 1996; Xu, 1999; Yuan, 2004, among others). Under corrective focus, to convey different degrees of emphasis, however, speakers of Standard Chinese rely more consistently on duration; and F_0 manipulation seems to be more restricted, findings that are

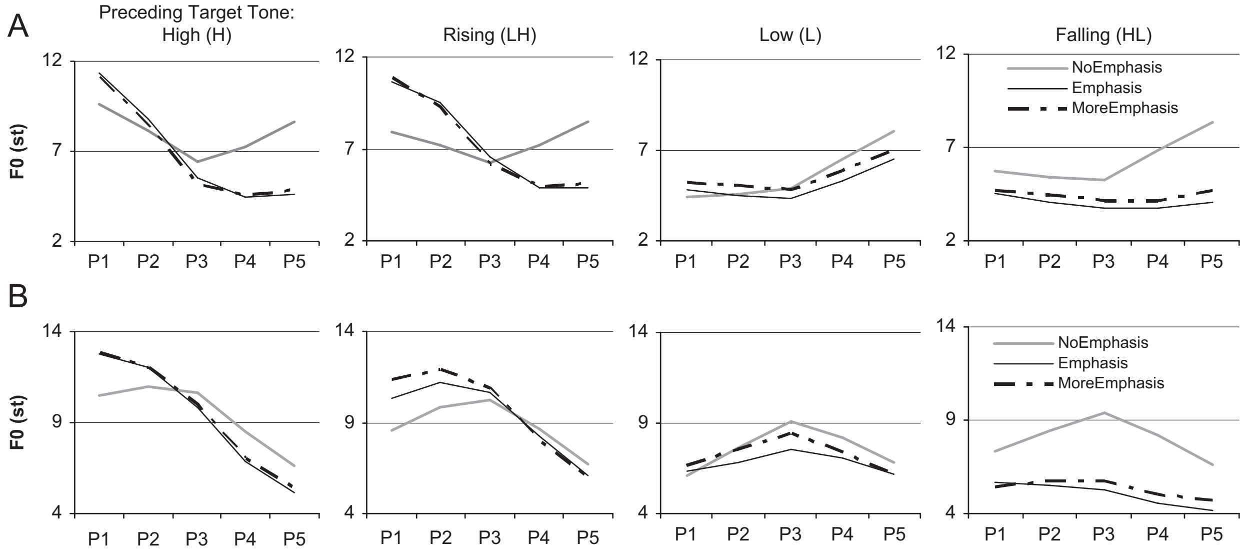

Fig. 13. Mean F_0 values of five equi-distant points over the Following syllable (Rising in A and Falling in B), as a function of different emphasis conditions and different Target tones (High in Column (C) 1; Rising in C2; Low in C3, and Falling in C4).

not consistent with what has been observed in English (Liberman & Pierrehumbert, 1984; Arvaniti & Garding, 2007).

The asymmetrical effect of emphasis on the durational and F_0 range adjustment, and, in particular, the fact that durational increase is a robust and consistent cue for degrees of emphasis in Standard Chinese, lend some support to the general hypothesis of functional load (Berstein, 1979) which states that the weighting of different acoustic cues for a linguistic function may be dependent on the functional load of the cues in conveying other linguistic information. In Standard Chinese, F_0 variation indicates lexical tonal contrasts and therefore is somewhat more restricted in conveying degrees of emphasis.

The expansion of the F_0 range was mainly due to the raising of F_0 maximum; F_0 minimum showed only a tendency of lowering. Given the serious creakiness in the emphasized Low tone and that creakiness often results from the lowering of F_0, it is important to note that in our data, emphasis certainly had an effect on the Low tone realization, although its phonetic realization cannot be assessed using F_0. Our data thus confirm, indirectly, prior studies that the change in F_0-range in Standard Chinese is better characterized as F_0-span expansion rather than F_0-level raising.

4.2. Testing the models of tonal implementation

In Section 1, we have identified a number of aspects of tonal realization that the three models of our interest make different predictions about: (1) F_0 range expansion, (2) tonal contour realization in terms of temporal alignment of minF_0 and maxF_0 as well as F_0 movements, and (3) the effect of the target tone on the F_0 realization of the preceding and following tones. The robust but asymmetrical changes in duration and F_0 as a function of emphasis now allow us to adjudicate between these three models.

With regard to F_0 range, we found a significant expansion from the NoEmphasis to the Emphasis condition, but no significant difference was found between the Emphasis and the MoreEmphasis condition. Moreover, lexical tones exhibited different limits on the extent to which their F_0 range could be expanded, with HL expanding more than LH and H. This tone-intrinsic variation in F_0 range was mainly due to variation of the maxF_0 (Fig. 6) (LH: 9.9 st for Emphasis and 11.0 st for MoreEmphasis; H: 13.1 st for Emphasis and 13.4 st for MoreEmphasis; HL: 13.7 st for Emphasis and 14.4 st for MoreEmphasis).[2]

[2] In the case of LH followed by HL, the F_0 peak over the following HL tone-bearing syllable was usually higher than that within the LH tone-bearing syllable in the two emphasis conditions. One possibility is that the maxF_0 of the emphasized LH tone was realized over the

All three models can account for a ceiling effect of F_0 range expansion. The Static-Target and Stem-ML models, without introducing extra mechanisms, would probably attribute such a ceiling effect to the speakers' physiological limits of maxF_0 raising or minF_0 lowering. However, given that maxF_0 values vary across tones, the lack of a robust increase in the case of H and LH cannot be attributed simply to physiological limits. This is particularly evident in the case of H when it was preceded by another H (Fig. 3). There should be enough time to realize the second H tone with considerably higher maxF_0 given that we know MaxF_0 could be higher (as in HL; see also Fig. 7) and also the speed of rise could be faster (as evident in the rise of H preceded by a Low tone). Contrary to our expectation, there was no significant increase in the maxF_0 of the high tone as the degree of emphasis increased. Instead, it was just the rising slope towards H that was adjusted according to whether H or L preceded the target H tone. This pattern replicates the results of Xu (1999). A similar pattern has also been observed in Spanish (Prieto, van Santen, & Hirschberg, 1995) and Greek (Arvaniti, Ladd, & Mennen, 1998).

PENTA offers a different explanation for the limitation of the pitch range. Focus is an independent pragmatic function in this model, encoded by means of a specific F_0 interval (as the non-local pitch target) within which the local pitch targets are implemented. This predicts that, despite the different degrees of emphasis with which corrective focus may be pronounced, the available F_0 range should remain the same. While this explains the non-significant expansion of the F_0 range from the Emphasis to the MoreEmphasis condition, it does not explain the tone-specific findings, unless different lexical tones are specified with different F_0 ranges for focus in different tonal contexts.

With regard to the F_0 contours, we found that the start of rise in LH and the fall in HL was significantly delayed, and correlated well with the duration of the tone-bearing syllable (Fig. 8). The start of these movements was also sensitive to the preceding tone (Fig. 9). The fall of HL started significantly earlier when preceded by H than by L, presumably because it took extra time for F_0 to rise from the low ending of L to the high beginning of HL. For LH, however, there was no significant effect of the preceding tone on the start of rise, suggesting controlled behaviour by speakers in delaying the start of rise, particularly when the preceding tone was L. In such cases, a sustained low pitch was produced, followed by a clear F_0 rise.

We also observed a significant increase in the absolute distance between the F_0 minimum and maximum in the case of LH and HL (Fig. 10) suggesting that the F_0 movement gestures were magnified. Their relative distance, however, only showed a tendency to increase with emphasis. There was also a significant increase in the steepness of the slope between the NoEmphasis and Emphasis conditions (Fig. 11). In short, delayed start of the rise/fall, magnified tonal gestures, and increased speed of the rise/fall all contributed to the distinctive realizations of LH and HL in the emphasis conditions, as shown in Fig. 3.

All three models are compatible with some, though not all of the above observations. Within the StaticTarget model, we expect that as the syllable duration increases, the absolute distance between minimum and maximum F_0 should increase correspondingly, whereas that their relative distance should remain unchanged. This idea is consistent with the finding that the slope of the movement seemed to be just a by-product of the locations of the static targets, and was given by the F_0 interpolation between them. Furthermore, we expect that the low and high targets of the individual tones of LH and HL should be temporally aligned with specific points in the segmental string, with some durational interval from those points, which, however, was not born out especially given the different alignment patterns of the HL and LH tones.

PENTA does not predict that the absolute distance between F_0 minimum and maximum should increase, while the relative distance should remain constant. In fact, the opposite pattern would appear the more reasonable prediction, as the model assumes F_0 movements, rather than F_0 turning points, as the primitives of tonal contours (for similar view on the dynamic nature of tonal targets in intonational languages, see, e.g. Ashby, 1978; 't Hart, Collier, & Cohen, 1990). Specifically, the dynamic tonal targets of LH and HL should be expected to exhibit rather constant rise/fall sizes (i.e. the absolute distance between the F_0 minimum and

(*footnote continued*)
following syllable, as one of the reviewers pointed out. We therefore also measured the maxF_0 of the following HL tone when the target LH tone was emphasized. The average maxF_0 over the HL tone-bearing syllable was 11.7 st for Emphasis and 12.3 st for MoreEmphasis without statistical significance between the two conditions. Both values were still much lower than the maxF_0 of HL tone in the respective emphasis conditions.

maximum) or rate of change (i.e. the rising or falling slope). PENTA would have to make explicit that when a dynamic tone is to be realized with a specific F_0 range (for emphasis), both the slope and the size of the movement are to be adjusted.

Neither StaticTarget nor PENTA, however, is able to capture the essence of the very striking pattern of tonal realization under emphasis: the distinctive realization of the lexical tonal contours. Stem-ML can model this more effectively, because it predicts that tonal gestures under emphasis should be magnified and realized with more distinctive versions of the lexical tones' characteristic F_0 contours. The increased distance between F_0 minimum and maximum and the stable relative distance are not surprising under the assumptions of this model. Also, the steeper slopes observed under corrective focus are the expected outcome of canonical realizations of rising and falling contours.

A challenge for all models, however, is the speakers' controlled realization of the start of rises and falls, in particular the finding that the rise for LH was delayed when preceded by L. A possible explanation of the delayed start of rise is perceptual salience. House (1999) shows that listeners are better able to judge pitch changes when the F_0 movement is present in the vowel than in the onset consonant. This, however, would leave the relatively earlier start of HL unaccounted for. Another possible explanation, along the same line of perceptual salience, is based on the need for maximal distinction within the tonal inventory. When there is a preceding L, both H and LH are realized with a rising F_0 contour, as illustrated in Fig. 3. To be maximally differentiated from H, the start of the rise would need to be delayed in LH. It could be argued that an early start of fall for HL could also potentially cause confusion with L when both are preceded by a H. Note, however, that HL is sufficiently differentiated from L by means of the slight rise before the fall, as shown in Fig. 3. A LH, however, cannot easily lower further after a L. Comparable context-sensitive adjustments in the phonetic implementation of tonal contrasts have also been reported by Smiljanić (2004) who observes that in Serbian, due to the presence of lexical tone contrast, the modification of the tones under focus is more constrained than in Zagreb Croatian, which lacks a lexical tone contrast.

Turning now to the effect of discourse context on the co-articulation of the target, preceding and following tones, Fig. 12 shows a marked asymmetrical pattern. Specifically, there was a more robust effect of the target lexical tone and discourse context on the onset F_0 of the following tone, than that on the offset F_0 of the preceding tone. This counters the prediction of Stem-ML and Static-Target, but follows precisely that of PENTA. Furthermore, the effect of discourse context and tonal coarticulation lasted throughout the whole following syllable (Fig. 13). A particularly interesting observation, which is not predicted by any of the models, is that post-focus tonal realization does not necessarily show a compressed F_0 range in certain tonal contexts (e.g. post-focus HL tone after a High tone). Rather, a more consistent description of post-focus tonal realization is the lack of sharp or precise F_0 contours that are characteristic the lexical tones.

4.3. Conclusion and implications

It is clear that all three models were capable of explaining some aspects of the tonal realization under degrees of emphasis, but failed in others. To summarize, our data suggest that an adequate model would need to account for the following three observations. First, there was a robust gradual increase in syllable duration from the NoEmphasis, via the Emphasis to the MoreEmphasis condition. The F_0 range expansion, however, was non-gradual: although there was a substantial increase from the NoEmphasis to the Emphasis condition, the expansion from the Emphasis to the MoreEmphasis condition was of a much smaller magnitude and also was not statistically significant. Furthermore, the expansion of F_0 range was tone-intrinsic in that each tone exhibited a different maxF_0, which argues against the possibility that the lack of robust and consistent F_0 range expansion from the Emphasis to the MoreEmphasis condition may be due to the ceiling effect of speakers' physiological limits. The tone-intrinsic F_0 range expansion also raises doubts about the position that pragmatic functions like focus are produced with a specific F_0 range with which the lexical tones are to be implemented.

Second, across the three emphasis conditions, the target tone had a significant and robust effect on the F_0 contour of the following tone, especially during the earlier portion of the tone-bearing syllable. This is in sharp contrast to the negligible effect of the target tone on the preceding tone. It suggests that the implementation of the lexical tones, regardless of emphasis condition, is better accounted for as continuous approximation of their pitch targets throughout the tone-bearing syllable, as proposed in Xu and Wang (2001).

Together with the specific details of the F_0 range expansion and syllable duration increase, a number of further phonetic adjustments under emphasis make it clear that lexical tones are produced with enhanced distinctiveness of their F_0 contours. Specifically, the wider F_0 range was mainly manifested by raised F_0 maxima, and sometimes by lowered minima, depending on the identity of the tone. Moreover, the low F_0 of an L tone was sometimes cued by creakiness. In the case of LH and HL tones, they showed clear rising or falling F_0 contours despite the significant increase of syllable duration. Moreover, the rise for the LH tone was strongly delayed, while the HL tone had a delayed as well as raised F_0 peak. As an *ensemble*, these adjustments enhance the distinctiveness of the contrasts among the lexical tones. Observing that emphasized lexical tones are realized with a larger F_0 range and longer duration grossly underreports the findings of this paper.

The effects of emphasis conditions on both the F_0 range and F_0 contour realization jointly suggest that focus cannot be a simple function of pitch range manipulation. Rather, tonal realization under emphasis seems analogous to vowel articulation under emphasis. A large number of studies have shown that vowels in prosodically prominent positions are longer and vowel targets are hyperarticulated (Fourakis, 1991; Lindblom, 1963; Moon & Lindblom, 1994). The English vowel [a] (in *plam*), for example, is usually produced with higher F_1 and lower F_2, such that it is realized lower and further back. As a contrast, [i] (in *seem*) is often produced with lower F_1 and higher F_2, such that it is realized higher and further front. A variety of observations have been made that express such observation as greater articulatory precision (Gussenhoven, 2004), the expansion of vowel sonority (Beckman, Edwards, & Fletcher, 1992), hyperarticulated phonemic features (Cho, 2005; de Jong, 1995), or more generally an increase in the contrastiveness of vowel quality (Erickson, 2002). In particular this latter characterization applies well to what we have observed about the effect of emphasis on tonal realization.

We tentatively propose that the difference between the NoEmphasis and Emphasis conditions is discrete, while that between the Emphasis and MoreEmphasis conditions is gradient. This would seem an appropriate interpretation both of the phonetic evidence that we have presented and from a functional point of view. The first difference reflects a difference between ‘out of focus’ and ‘with corrective focus’, while the second is concerned with different degrees of articulatory force for one of these conditions. This assumption raises the question of what representational difference should exist between the NoEmphasis condition on the one hand and the Emphasis and MoreEmphasis conditions on the other. Two commonly encountered options would appear to be out of the question. First, the pitch accentuation of focus constituents and the deaccentuation of out-of-focus constituents of the sort proposed for many languages, most notably West Germanic, is inapplicable, since both the NoEmphasis and the two emphatic conditions had the same tonal representation, as the lexical tones were held constant. Second, the occurrence of a prosodic phrase break on the left or the right of the focus constituent, as reported for a variety of languages (e.g. Jun, 1993; Pierrehumbert & Beckman, 1988; Selkirk & Shen, 1990, among others) could not obviously be ascertained, since Third-tone sandhi applied to the experimental syllable in all conditions, suggesting that the phrasing remained constant.[3] Moreover, it is not obvious that these representations would capture the important finding reported here that focus affects the pronunciation at the segmental level and at the suprasegmental level in comparable ways.

One way in which the effect of focus at both the segmental and suprasegmental levels might be accounted for is by appealing to an abstract notion of metrical prominence, along the line of Truckenbrodt (1995), Ladd (1996), Selkirk (2002) and Féry and Samek-Lodovici (2005) (among others). This would mean that the corrective focus constituent in Standard Chinese has the highest prosodic prominence of the utterance, i.e. is associated with the strongest node in a metrical tree. In English, this representation leads to the association of nuclear pitch accent with the contrastively focused constituent (Selkirk, 2002), causing that syllable to be articulated with greater articulatory force (Cho, 2005; de Jong, 1995; Erickson, 2002). In languages in which the focus constituent obeys an edge constraint, as have been claimed for Japanese, the highest metrical level leads to the coincidence of the focus constituent with a phonological boundary, causing the focus constituent to have wider pitch range, possibly in addition to greater articulatory force (Pierrehumbert & Beckman, 1988,

[3] This observation is surprising in view of Shih (1997) and Zhang (1988), both of which report a blocking effect of emphasis on the Third-tone sandhi. All five subjects in this study, however, consistently applied the tone sandhi rule in all three emphasis conditions. As one of the reviewers pointed out, it is possible that although they also used the term emphasis, Shih (1997) and Zhang (1988) have probably elicited a different type of focus.

but also see Ishihara, 2005). In Mandarin Chinese, where there is no addition of a pitch accent or edge requirement, the greater articulatory force applies equally to segments and tones of the focused syllables.

A metrical representation giving the focus the highest level of prominence mediates between the focus marking and the phonetic realization, and makes it unnecessary to assume that the focus marking is directly interpreted by the phonetic interpretation component. While metrical prominence may provide an account for the prosodic realization of focus as far as the lexical tones in Standard Chinese are concerned, and in that way may unify the prosodic effect of focus in tone languages like Standard Chinese, the distribution of focus-marking pitch-accents in languages like English, and the edge-based focus effects in other languages, the question arises to what extent metrical prominence can explain the prosodic realization of different types of focus, or how the prosodic encoding of information structure notions like topic can be incorporated. Also, in languages in which the focus constituent does *not* have the highest level of prominence, like Northern Sotho, Yucatec Maya and Chitumbuka (Downing, 2008; Gussenhoven & Teeuw, 2007; Zerbian, 2006), the requirement that the focus constituent has the highest level of prominence must be outranked by constraints that require prominence to be elsewhere, such as lengthening of the phrase-penultimate vowel independently of the location of the focus constituent in Northern Sotho and Chitumbuka.

Acknowledgements

Questions and comments from Ellen Broselow, Marie Huffman, Chilin Shih, Yi Xu and Jiahong Yuan on earlier write-ups are gratefully acknowledged. We are also grateful to the comments of two anonymous reviewers and the editor Jonathon Harrington on an earlier version, and to all our speakers for their participation. Usual disclaimers apply. The partial support by the VENI grant from the Netherlands Organization for Scientific Research (NWO) to the first author is gratefully acknowledged.

References

Arvaniti, A., & Garding, G. (2007). Dialectal variation in the rising accents of American English. In J. Cole, & J. Hualde (Eds.), *Laboratory phonology* (Vol. 9, pp. 547–576).

Arvaniti, A., Ladd, R., & Mennen, I. (1998). Stability of tonal alignment: The case of Greek prenuclear accents. *Journal of Phonetics*, *26*(1), 3–25.

Ashby, M. (1978). A study of two English nuclear tones. *Language and Speech*, *21*, 326–336.

Beckman, M., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In J. Docherty, & R. Ladd (Eds.), *Papers in laboratory phonology II: Gesture, segment, prosody* (pp. 68–86). Cambridge: Cambridge University Press.

Berstein, A. E. (1979). *A cross-linguistic study on the perceptual and production of stress*. Ph.D. dissertation, Linguistics Department, University of California at Los Angeles, Los Angeles.

Boersma, P., & Weenink, D. (2005). *Praat: Doing phonetics by computer*. (version 4.3.14) [Computer program]. Available at: 〈http://www.praat.org/〉.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chen, Y. (2003). *The phonetics and phonology of contrastive focus in standard Chinese*. Ph.D. dissertation, Stony Brook University, Stony Brook.

Chen, Y. (2006). Emphasis, syllable duration, and tonal realization in Standard Chinese. In *Speech prosody 2006*. Dresden.

Chen, Y., & Braun, B. (2006). The prosodic realization of information structure categories in Standard Chinese. In *Speech prosody 2006*. Dresden, NY.

Chen, Y., & Xu, Y. (2006). Production of weak elements: Evidence from neutral tone in Standard Chinese. *Phonetica*, *63*, 47–75.

Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a, i/ in English. *Journal of the Acoustical Society of America*, *117*(6), 3867–3878.

de Jong, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *Journal of the Acoustical Society of America*, *97*(1), 491–504.

Downing, L. (2008). Focus and prominence in Chichewa, Chitumbuka and Durban Zulu. *ZAS Papers in Linguistics*, *49*, 47–65.

Duanmu, S. (2000). *The Phonology of Standard Chinese*. Oxford: Oxford University Press.

Erickson, D. (2002). Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, *59*, 134–149.

Féry, C., & Samek-Lodovici, V. (2005). Focus projection and prosodic prominence in nested foci. *Language*, *82*(1), 131–150.

Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America*, *90*, 1816–1827.

Fujisaki, H. (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In O. Fujimura (Ed.), *Vocal Physiology: Voice Production, Mechanisms and Functions* (pp. 347–355). New York: Raven.

Gårding, E., Zhang, J., & Svantesson, J. O. (1983). *A generative model for tone and intonation in Standard Chinese based on data from one speaker*. Lund University Working Papers 25 (pp. 53–65).

Gårding, E. (1979). Sentence intonation in Swedish. *Phonetica*, *36*, 207–215.

Grønnum, N. (1995). Superposition and subordination in intonation: a non-linear approach. *Proceedings of the 13th International Congress of Phonetic Sciences*, (pp. 124-131).

Gussenhoven, C. (1983). Testing the reality of focus domains. *Language and Speech*, *26*(1), 61–80.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Gussenhoven, C., & Teeuw, T. (2007). A moraic and a syllabic H-tone in Yucatec Maya. In E. Herrea Z., & P. M. Butrageño (Eds.), *Fonología instrumental: Patrones fónicos y variacion lingüística* (pp. 49–71). Mexico City: Colegio de México.

Gussenhoven, C., & Rietveld, A. C. M. (2000). The behavior of H* and L* under variations in pitch range in Dutch rising contours. *Language and Speech*, *43*, 183–203.

House, D. (1999). Perception of pitch and tonal timing: implications for mechanisms of tonogenesis. In *Proceedings of ICPhS-99* (pp. 1823–1826).

Ishihara, S. (2005). Prosody-scope match and mismatch in Tokyo Japanese Wh-Questions. *English Linguistics*, *22*(2), 347–379.

Jin, S. (1996). *An acoustic study of sentence stress in Mandarin Chinese*. Ph.D. dissertation, The Ohio State University, Columbus, OH.

Jun, S.-A. (1993). *The phonetics and phonology of Korean prosody*. Ph.D. dissertation, The Ohio State University, Columbus, OH.

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

Ladd, D. R. (1980). *The structure of intonational meaning*. Bloomington: Indiana University Press.

Ladd, R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics*, *25*(3), 313–342.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff, & R. Oehrle (Eds.), *Language, sound, structure: Studies in phonology presented to Morris Halle by his teacher and students* (pp. 157–233).

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*, 1773–1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In M. Hardcastle (Ed.), *Speech production and modeling*. Dordrecht: Kluwer Academic Publishers.

Moon, S., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America*, *96*, 40–55.

Pierrehumbert, J., & Beckman, M. (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Prieto, P., van Santen, J., & Hirschberg, J. (1995). Tonal alignment patterns in Spanish. *Journal of Phonetics*, *23*(4), 429–451.

Rietveld, A. C. M., & Gussenhoven, C. (1985). On the relation between pitch excursion size and pitch prominence. *Journal of Phonetics*, *13*, 299–308.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge: MIT Press.

Selkirk, E. (2002). Contrastive FOCUS vs. presentational focus: Prosodic evidence from right node raising in English. In *Speech prosody 2002: Proceedings of the 1st international conference on speech prosody* (pp. 643–646), Aix-en-Provence.

Selkirk, E., & Shen, T. (1990). Prosodic domains in Shanghai Chinese. In S. Inkelas, & D. Zec (Eds.), *The phonology–syntax connection*. Chicago: University of Chicago Press.

Shen, X. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics*, *18*, 281–295.

Shih, C. (1988). Tone and intonation in Mandarin. *Working Papers of the Cornell Phonetics Laboratory*, *3*, 83–109.

Shih, C. (1997). Mandarin Third Tone sandhi and prosodic structure. In J. Wang, & N. Smith (Eds.), *Studies in Chinese phonology* (pp. 81–123). Mouton de Gruyter.

Sluijter, A. (1995). *Phonetic Correlates of Stress and Accent*. The Hague: Holland Academic Graphics.

Smiljanić, R. (2004). *Lexical, Pragmatic, and Positional Effects on Prosody in Two Dialects of Croatian and Serbian: An Acoustic Study*. New York: Routledge.

Truckenbrodt, H. (1995). *Phonological phrases: Their relation to syntax, focus and prominence*. Ph.D. dissertation, MIT, Cambridge, MA.

’t Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach*. Cambridge: Cambridge University Press.

Turk, A., & Sawusch, R. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics*, *25*, 25–41.

Wu, Z. (1982). Rules of intonation in Standard Chinese. *Preprints from the 13th International Congress of Linguistics* (pp. 95–108).

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*(1), 61–83.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F_0 contours. *Journal of Phonetics*, *27*(1), 55–105.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*, 220–251.

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *JASA*, *111*, 1399–1413.

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

Xu, Y., & Xu, X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*, *33*, 159–197.

Yuan, J. (2004). *Intonation in Mandarin Chinese: Acoustic, perception, and computational modeling*. Ph.D. dissertation, Cornell University, Ithaca, NY.

Zerbian, S. (2006). *Expression of information structure in the Bantu Language Northern Sotho*. Ph.D. dissertation, Humboldt-Universität zu Berlin.

Zhang, Z. (1988). *Tone and Tone Sandhi in Chinese*. Ph.D. dissertation, The Ohio State University, Columbus, OH.