Journal of Speech Sciences 3(1):85-140.2013
Available at: hhtp://www.journalofspeechsciences.org

Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling

Fang Liu[1], Yi Xu[2], Santitham Prom-on[2,3], Alan C. L. Yu[4]

[1] Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong, China.[2] Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, UK.

[3] Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand

[4] Department of Linguistics, University of Chicago, 1010 E. 59th St., Chicago, IL 60637, USA.

Abstract

In this paper we address the long-standing issue of how prosodic patterns are linked to meanings. We explore the idea that prosodically realized communicative functions, such as focus and sentence modality, are analogous to lexical morphemes, the smallest sound units that carry meaning. We considered evidence of a four-way similarity between lexical morphemes and prosodic functions. First, similar to lexical morphemes, each prosodic function consists of multiple phonetic components. Second, like segmental phonemes, individual prosodic components are meaningless themselves, but act jointly to mark both intra- and inter-functional contrasts. Third, like lexical morphemes, prosodic functions have allomorph-like variants whose occurrences are conditioned by factors like location in sentence and interaction with other prosodic functions. Finally, similar to lexical morphemes, prosodic functions are language-specific and the specificity has likely historical sources. We examined the evidence by a) reviewing existing literature on speech prosody, b) conducting two new experiments on the production of focus and sentence modality in General American English and Mandarin Chinese, and c) training an articulatory-functional model on focus, modality, tone and stress in English and Mandarin, and synthesizing fully-detailed F_0 contours with the learned functional targets. Overall, all the evidence examined is in support of the hypothesis. In particular, the consistency between the target parameters obtained from acoustic analysis and computational modeling, and the close match between functionally synthesized and naturally produced F_0 contours demonstrate the plausibility of establishing a clear link between function-specific categorical representations and fine-detailed surface prosody.

Keywords: Morpheme; Prosodic functions; Focus; Sentence modality; Mandarin; English

Corresponding author: Yi Xu, Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, UK. E-mail: yi.xu@ucl.ac.uk

“The linguist’s theory of intonational meaning is extremely widespread among linguists of otherwise diverse outlooks.”

> “The central idea of this view is that *the elements of intonation have morpheme-like meaning*.”
>
> “This view contrasts sharply with the assumptions underlying the instrumental approach, in which it is generally assumed that quite specific meanings, such as interrogation, anger, and incompleteness, are conveyed by rather general phonetic properties, such as overall raising of pitch; and that context-dependent pragmatic inference plays little role in the interpretation of intonational features.” (Ladd, 2008:41; original emphasis) (1)

1. Introduction

How prosodic forms are linked to meaning is a long-standing issue, and numerous proposals have been offered (1-10). As summarized by Ladd (2008) in the above quotes, at least two diametrically different approaches can be identified. (1) One assumes that there exist “elements of intonation” that carry “morpheme-like meaning,” hence are directly meaningful, where the elements are units like pitch accents, phrase accents and boundary tones that are at the same time also phonological. (1) There are different versions of this approach and so far no clear consensus has been reached. As pointed out by Ladd (2008), “there is no theoretical framework within which we can undertake a comparative evaluation that would command general agreement… For the present, proposals about intonational meaning are not a reliable source of evidence on intonational phonology.” The other approach, referred to by Ladd (2008) as being instrumental, tries to empirically identify prosodic correlates of meaningful functions such as focus and interrogation, and many have been found. (12-17) This approach, however, often gives the impression that any conceivable communicative function should have clear phonetic correlates, irrespective of linguistic factors or language differences. This is becoming increasingly incompatible with findings of language specificity in prosody. (18-19) In general, therefore, how exactly prosody is linked to meaning remains unclear.

The present paper is an attempt to clarify the issue of prosody-meaning link by combining the insights of both linguistic and instrumental approaches, but to do so on a solid empirical basis. We start by first considering the assumption of the “linguist’s theory” that prosody consists of morpheme-like structures that are directly meaningful. This assumption makes good theoretical sense because, at the lexical level, it is a fundamental notion that the meaning-sound link is achieved through morphology, i.e., semantic meanings are carried by morphemes rather than segments or syllables directly. Specifically, lexical morphemes are defined as the smallest units that carry meaning. (11) For example, the word *uncertainty* consists of four syllables but only three morphemes, *un-*, *certain*, and *-ty*. None of the morphemes can be broken up further without losing its meaning. Not only cannot the monosyllabic morpheme ‘-ty’ be broken up into /t/ and /i/, but also the disyllabic morpheme *certain* cannot be broken up into "cer" and "tain," because neither the phonemes nor the syllables represent meanings on their own. There are also other recognizable properties of lexical morphemes. Together they can be summarized as follows:

1. Non-autonomy of components — components of a morpheme cannot be separated from each other while still conveying the same meaning on their own, unless they have already been developed into allomorphs. This is the most essential property of morpheme because it is directly derived from its definition, i.e., being the minimal meaningful unit.

2. Multi-componential coding — a morpheme may consist of variable number of phonological components, from a single phoneme, e.g., the English plural ‘s’, to multiple syllables, e.g., /mai4ke4feng1/ in Mandarin (borrowed from the English *microphone*, where the numerals mark the tones).

3. Conditional allomorphs — a morpheme often has allophone-like alternative forms, whose occurrences are conditioned by recognizable factors. A well-known example is the plural form ‘s’ in English, which varies from /s/ in *cats*, /z/ in *dogs*, to /∂z/ in *dishes*.

4. Language-specificity with diachronic sources — each lexical morpheme has its own etymology. For example, *Karaoke* was borrowed into English and many other languages from Japanese only in the 1980s, while *mother* in English likely has a very long etymology with multiple historical changes.

Given these properties, it would be possible to examine whether they could also be found in the prosodic domain. Importantly, however, the examination needs to be empirical, like what is usually done in the “instrumental approach.” (1) In the rest of this paper, we will first examine empirical evidence for morpheme-like prosodic structures with systematic acoustic analysis of English and Mandarin intonation as well as computational modeling of the same structures in the two languages. As we will show, there is indeed evidence of prosodic structures that exhibit characteristics of lexical morphemes. However, it is the holistic communicative functions such as focus and modality (statement vs. question), rather than individual units such as pitch accents, phrase tones, or boundary tones, that seem to bear the most similarities to lexical morphemes. Finally, we will review existing literature for further evidence, especially in regard to language-specificity and historical linearity of prosodic structures.

2. Experiment 1 — Acoustic analysis of focus and question intonation in English

This experiment examines evidence of morpheme-like properties in English in two hypothetical communicative functions, *focus* and *modality*. Questions about morpheme-like properties have not been specifically asked in previous research, but many issues addressed before are nevertheless relevant. The first is about the temporal scope of the acoustic manifestations of focus and modality, which is related to the property of *multi-componential coding*. For both focus and modality, there is evidence that their marking involves not just a single acoustic dimension such as F_0, or just a single location in a sentence. Focus involves both on-focus increase of F_0, duration, intensity, and upper spectral energy, and post-focus reduction of F_0 (12-16). Question marking involves not only sentence-final F_0 contours, as has been long recognized, (17, 20-22) but also acoustic patterns that occur as early as the first word bearing sentence stress. (17, 22)

The second issue is the interaction between modality and focus, which is related to the property of *conditional allomorphs*. As shown by Eady and Cooper (1986) and Pell (2001), in sentences with initial focus, maximum F_0 of statements and questions differs mainly in post-focus words: dropping to a low level in statements but remaining high in questions. (14, 17) In sentences with final focus, statements and questions differ mainly in maximum F_0 of the final focused word: higher in questions than in statements. Furthermore, regardless of focus condition, focused words exhibit a rising F_0 contour in questions but a falling F_0 contour in statements. (17) Nevertheless, many details about focus and modality and their interaction with each other remain unclear for English. Eady and Cooper (1986) and Pell (2001) reported only F_0 of entire words, but not of individual syllables, (14, 17) leaving the prosodic properties of many syllables unspecified. They also only compared peak F_0 values of a few key words, leaving the details of the rest of F_0 contours largely unexamined. Probably because of the coarseness of the analysis, no F_0 differences were reported between statement and question in sentences with neutral or sentence-final focus. This contrasts with the finding for Danish by Thorsen that the modality contrast becomes perceptible to native listeners after the first stress group. (24)

The difficulty with examining detailed F_0 contours is that it is hard to compare entire F_0 contours point by point. And the difficulty is further exacerbated by the vast amount of contextual variability that occurs even when the underlying tonal category is relatively fixed as in the case of lexical tone. (25-26) Nevertheless, recent tone language research has shown that the highly variable surface tonal contours can be linked to relatively invariant simple underlying pitch targets, and that much of the surface variability can be attributed to an articulatory process of syllable-synchronized target approximation, as characterized by the target approximation (TA) model, shown in Figure 1. The figure illustrates how simple linear underlying targets can lead to continuous surface F_0 contours, thus allowing the representation of complex surface F_0 contours in their entirety with simple targets that can be specified in terms of just height and slope. The adequacy of such representation is attested by computational simulations based on the quantitative target approximation (qTA) model, (27), which will be described in detail in Experiment 3. In the same study, it is shown that it is also possible to identify, for English, syllable-sized underlying pitch targets (28) that can generate F_0 contours that closely match those of natural utterances. Those identified targets are associated with lexical stress and focus, in agreement with early findings based on acoustic analysis. (17)

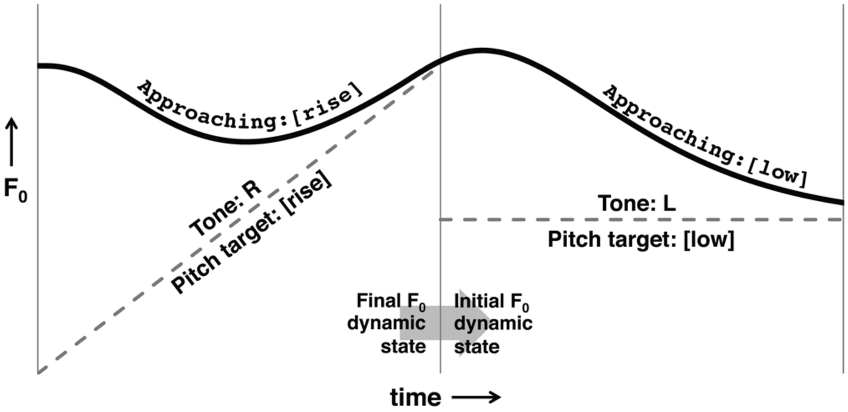


Figure 1. A sketch of the Target Approximation (TA) model. (28) The thick solid curve represents the F_0 contours that asymptotically approach two successive pitch targets represented by the dashed lines. The middle vertical gray line represents the syllable boundary through which the final-F_0 dynamic state is transferred from the first syllable to the second. Such state transfer results in a smooth and continuous F_0 trajectory across syllable boundaries.

The goal of Experiment 1 is therefore to go beyond what is known from previous studies of English intonation and establish all the function-specific underlying representations of focus and modality that will allow the explanation of the full details of surface F_0 contours in the language. These function-specific representations should also be consistent with the computational model parameters to be obtained in Experiment 3, which are capable of generating detailed F_0 contours that closely match those of the natural utterances. In particular, instead of taking measurements like maximum or mean F_0, which only indicate pitch range variations, we will take F_0 height and slope near the offset of each syllable, which would indicate both pitch range variations and properties of syllable-sized underlying pitch targets. The rationale for these two measurements can be seen in Figure 1, where it is clear that regardless of the initial transitional variations, the surface F_0 contours most closely approximate an underlying target by the end of the corresponding syllable in terms of both height and slope. Additionally, we will examine how duration and intensity also vary with focus and modality in English. More specifically, the following research questions will be examined for General American English:

1) Are there syllable-sized underlying pitch targets associated with focus and modality?

2) Are there intensity and durational differences also associated with focus and modality?

3) What are the temporal domains of focus and modality coding, respectively?

4) Are there any interactions between focus, modality and lexical stress in English?

Answers to these questions will help to establish whether there is evidence for *multi-componential coding* as well as *conditional allomorphs* of focus and modality in English.

2.1. Materials

Test materials consisted of three sets of sentences, within which the final syllable of the last word is either stressed (e.g., *E**laine***, ***May***) or unstressed (e.g., *A**lan***, ***me***), as shown below. The design is to enable the examination of underlying pitch targets under the interaction of focus, modality, lexical stress and location of syllable in word and sentence.

1. You want a **job** with **Microsoft**./?

   You want a **job** with **La Massage**./?

2. There is something **unmarriable** about **me**./?

   There is something **unmarriable** about **May**./?

3. You're going to **Bloomingdales** with **Alan**./?

   You're going to **Bloomingdales** with **Elaine**./?

Each sentence was produced as either a statement or yes/no question, and with focus on either the sentence-medial or sentence-final word. The modality and focus conditions were elicited by different prompt sentences, with which mini-dialogues were formed (see Appendix 1 for details). Such mini-dialogues created specific pragmatic contexts that form "context-dependent pragmatic inference." (1: page 41) All the sentences were repeated eight times by each subject in separate blocks, each with a different randomized order, resulting in 960 utterances in total.

2.2. Subjects

Three female and two male speakers, aged 18-30, participated as subjects. They were raised in either California or the Midwest in the United States, and spoke General American English. None reported having speech or hearing disorders.

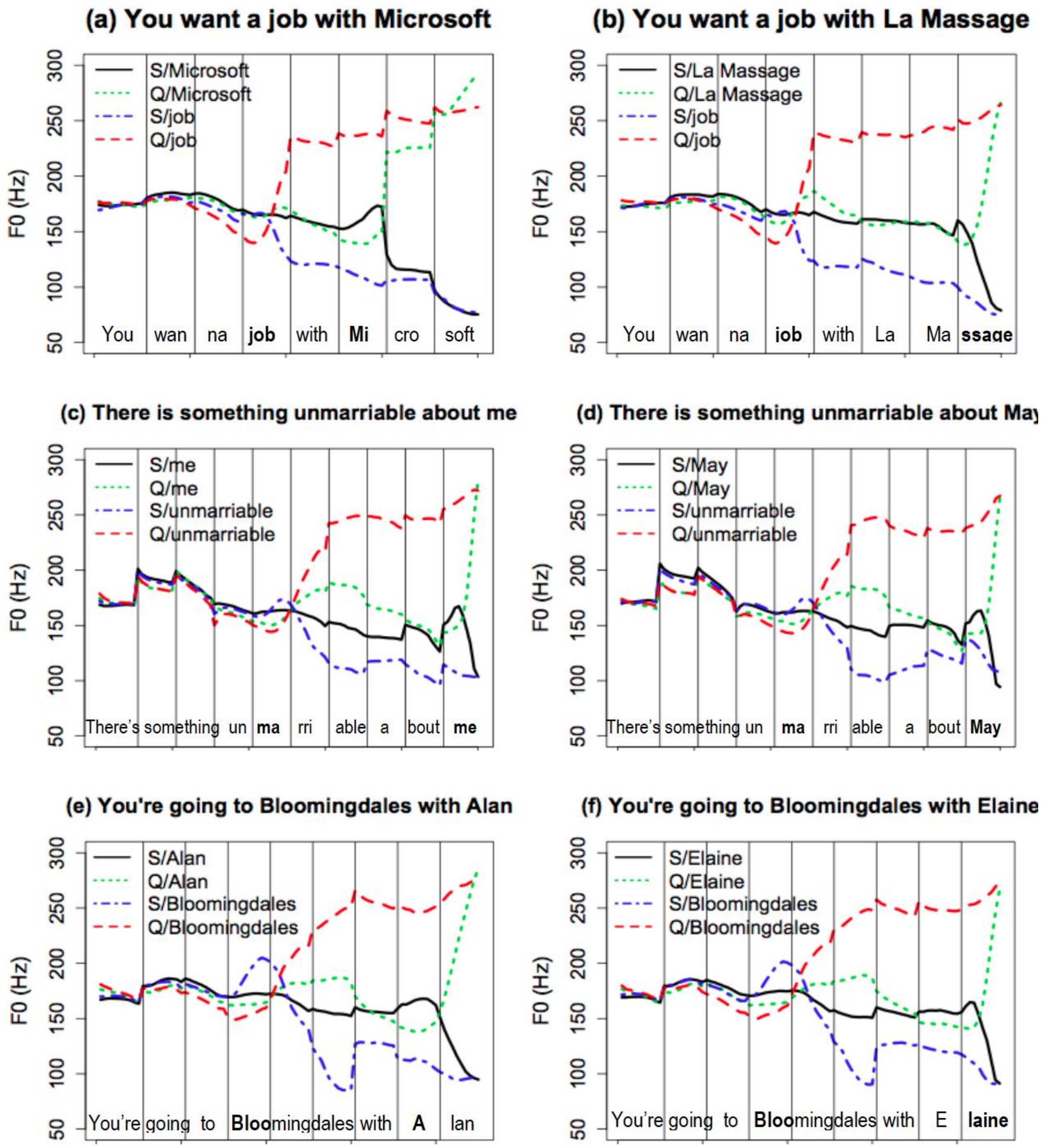
2.3. Procedure

Recordings were done in a sound-treated booth in the Language Labs at the University of Chicago, Chicago, Illinois. Subjects were first familiarized with the test materials before the start of the recording. During recording, the target sentences were displayed on a computer screen together with the corresponding prompt sentences, one mini-dialogue at a time. The subject read aloud both the prompt and target sentences[i], and the sounds were digitized at 22.05 kHz using a solid-state recorder.

For visual inspection and graphic analysis, a Praat (29) script computed time-normalized F_0 contours of the sentences by getting the same number of evenly spaced F_0 points from each syllable (see Xu (2005-2013) (30) for a general-purpose version of the script). This technique allows direct graphic comparison of continuous F_0 contours as opposed to the more common practice of comparing only single (e.g., 12, 14, 17) or double measurements (e.g., 31). In this common practice, readers can only guess what the rest of F_0 contours may look like. Time-normalized F_0 contours leave much less to guesswork. For statistical analyses, three measurements were taken by the script from the raw F_0 contours of the stressed syllables in all the key words: a) final-F_0 (in st, indicating target height), b) final-velocity (instantaneous rate of change of F_0, = $(F_{0i+1} - F_{0i-1}) / (t_{i+1} - t_{i-1})$, in st/s, indicating target slope) were taken at 30 ms before syllable offset (see (32-34) for justification of the 30 ms offset), and c) syllable duration (in ms). Here the measurement of final-F_0 and final-velocity is based on the principle of the target approximation model. As illustrated in Figure 1, a pitch target, defined in terms of both height and slope, is best approximated by the end of the syllable to which it is associated. Measurements taken at the end of a syllable would therefore best reflect its underlying pitch target. Note that the time-normalized mean F_0 contours, which is also characteristic of our previous studies, is in addition to, rather than in place of, measurements taken for the purpose of statistical analysis. But close examination of mean time-normalized F_0 contours, especially when they are plotted in overlaid graphs according to the controlled factors, like in Figures 2, 6 and 7, enables us to choose the measurements that maximally reflect the real differences.

2.4. Results

Figure 2 displays mean time-normalized F_0 contours (averaged across 40 repetitions by 5 subjects in logarithmic scale) of the test materials under different focus conditions and in different modalities (with syllable as the normalization domain). From these graphs, various effects of the controlled factors can be seen. Overall, the height and slope of F_0 in a syllable seem to be influenced by all three factors: focus, modality, and location of the stressed syllable in word and sentence. In the interest of space, analysis will only focus on the stressed syllables in sentence-medial and -final positions.


(a) You want a job with Microsoft

(b) You want a job with La Massage

(c) There is something unmarriable about me

(d) There is something unmarriable about May

(e) You're going to Bloomingdales with Alan

(f) You're going to Bloomingdales with Elaine

Figure 2. Time-normalized F_0 contours (averaged across 40 repetitions by 5 subjects in logarithmic scale) of test materials in English. Vertical lines indicate syllable boundaries. In the legend, "S" stands for statement and "Q" for yes/no question. "S/Microsoft" means a statement with focus on "Microsoft", and so on.

2.4.1. Analysis of final-F_0 and final-velocity

Figure 3 displays final-F_0 and final-velocity of the key syllables summarized in four bar graphs, each in either sentence-medial or sentence-final position. These graphs show the effects of modality, focus, and position of the stressed syllable in word and sentence. All the significant effects ($p < .05$), based on four 3-way repeated measures ANOVAs, each corresponding to a graph, are shown in Table 1[ii]. (In this and subsequent statistics tables, the non-significant cells are left blank.) As can be seen, the only significant main effects are those of modality. All the other significant effects are interactions. For final-F_0, the strongest effect is the 3-way interaction between all three factors, and the same 3-way interaction is also significant for final-velocity, indicating that the effects of focus and modality are especially strong toward the end of the sentence. The significant 2-way interactions indicate how the underlying pitch targets of the key syllables may have varied. The interaction of focus and modality on both measurements show that the slope of the target tends to be falling (with negative velocity and sometimes lower final-F_0) in statements but rising in questions, and the tendency not only occurs in the sentence-final location but also in the sentence-medial location, especially when the stressed syllable is word-final and on-focus.

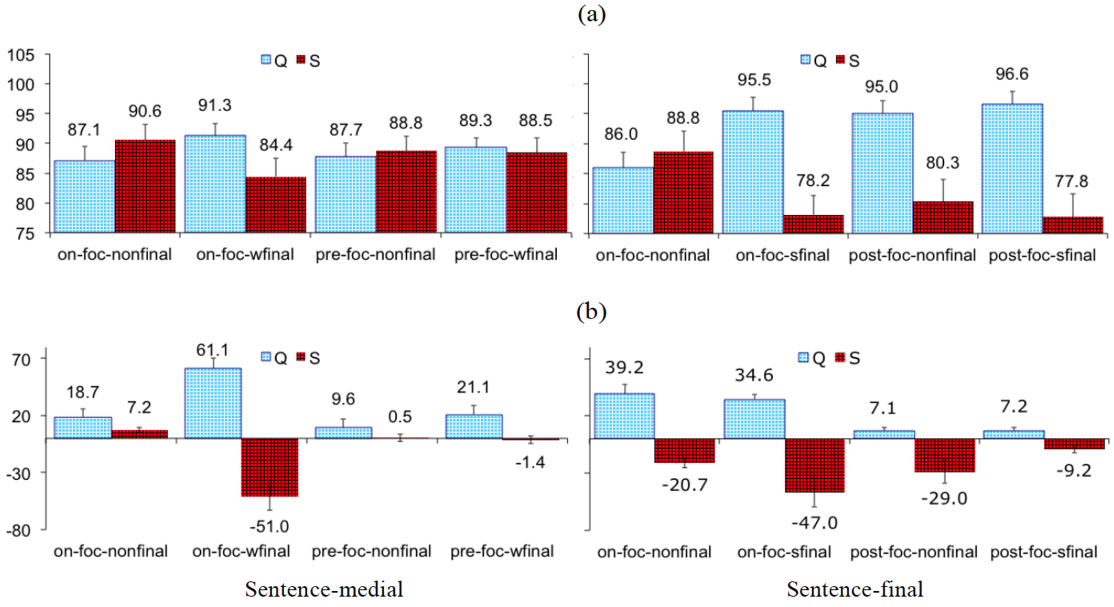


Figure 3. Final-F_0 and final-velocity values of the stressed syllables in the key words under different focus, modality, and syllable position conditions in English. Note: “foc” stands for focus, “wfinal” for word-final, and “sfinal” for sentence-final.

Table 1. Significant effects of modality, focus, and syllable position on (a) final-F_0 and (b) final-velocity of the stressed syllables in sentence-medial and -final positions.

	Significant effects	Sentence-medial		Sentence-final	
		F	*p*	*F*	*p*
Final-F_0 (*df* = 1, 4)	Modality			51.96	< .01
	Focus * Modality			34.60	< .01
	Modality * Syllable-position	26.40	< .01	159.68	< .001
	Focus * Modality * Syllable-position	15.98	< .05	260.08	< .001
Final-velocity (*df* = 1, 4)	Modality	34.30	< .01	36.33	< .01
	Focus * Modality	67.14	< .01	20.88	< .05
	Focus * Syllable position			13.84	< .05
	Modality * Syllable-position	63.39	< .01		
	Focus * Modality * Syllable-position	41.08	< .01	10.37	< .05

To further determine the pitch targets of the stressed syllables of the key words, *t*-tests were conducted to see if their final velocities are significantly different from zero, and the results are shown in Table 2. If final velocity is different from zero, the underlying target is likely to be either a rising (when the value is positive) or falling one (when the value is negative). The top half of the table shows the mean final velocities of the on-focus stressed syllables. The results indicate that the pitch targets of these syllables are likely to be high-level (non-word-final) or falling (word-final) in statements, but always rising in questions. The lower half of Table 2 shows the results of pre- and post-focus stressed syllables. Pre- and post-focus syllables are grouped together because their final velocities do not differ according to their position relative to focus (F(1, 3) = 5.69, *p* > .05). Results of *t*-tests indicate that the pitch target of the pre- or post-focus stressed syllable is level (if it is word-final but non-sentence-final) or falling (if it is non-final or word-final and sentence-final) in statements, but rising in questions. Note that the velocity values of the non-final stressed syllables in pre-/post-focus content words in statements are only marginally significantly different from zero (*t* = -2.20, *p* = 0.0357, Table 2), so they may have a level rather than falling target.

Table 2. Mean final velocities (st/s) of stressed syllables in English statements and questions. The t-tests indicate whether these velocities are significantly different from zero. Note: "w-final non-s-final" stands for word-final but non-sentence-final, and "w-final s-final" for word-final and sentence-final.

		Non-final	W-final, Non-s-final	W-final, S-final
On-focus	Question	29.70 t(29) = 5.95, *p* < .001	62.83 t(9) = 9.51, *p* < .001	41.00 t(19) = 12.72, *p* < .001
	Statement	-7.57 t(29) = -1.80, *p* > .05	-52.10 t(9) = -6.03, *p* < .001	-53.49 t(19) = -6.64, *p* < .001
Off-focus	Question	10.24 t(29) = 3.68, *p* < .001	17.72 t(9) = 2.47, *p* < .05	8.48 t(19) = 5.11, *p* < .001
	Statement	-9.63 t(29) = -2.20, *p* < .05	-2.05 t(9) = -0.66, *p* > .05	-15.96 t(19) = -4.38, *p* < .001

2.4.2. Intensity effects

Figure 4 displays mean intensity of the stressed syllables in the key words under different focus, modality, and syllable position conditions, and Table 3 shows the significant effects on mean intensity based on two 3-way ANOVAs. As expected, focused syllables have greater intensity than both pre- and post-focus syllables and the differences are highly significant. Intensity in questions is higher than in statements in sentence-final location. This is probably because, other things being equal, higher F_0 is associated with greater intensity. (35) The strong interaction between focus and modality at the sentence-final location indicates that post-focus reduction of intensity in statements, as found in previous studies, (13 and 19) is likely a byproduct of lowered post-focus F_0. In questions, as shown in Figure 4, post-focus intensity is no longer reduced. The strong interaction between modality and syllable position in the sentence-final location seems to be of the same nature when comparing Figure 4 with Figure 3a: the difference in intensity increases as the difference in F_0 increases.

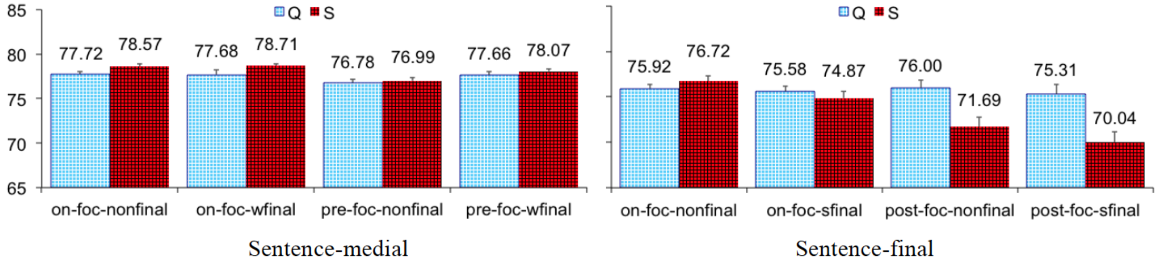


Figure 4. Mean intensity values of the stressed syllables in the key words under different focus, modality, and syllable position conditions in English. Note: "foc" stands for focus, "wfinal" for word-final, and "sfinal" for sentence-final.

Table 3. Significant effects of modality, focus, and syllable position on mean intensity of the stressed syllables in sentence-medial and -final positions.

Significant effects	Sentence-medial		Sentence-final	
(*df* = 1, 4)	*F*	*p*	*F*	*p*
Focus	33.35	< .01	23.26	< .01
Modality			19.41	< .05
Focus * Modality	12.97	< .05	144.31	< .001
Focus * Syllable position	28.92	< .01		
Modality * Syllable position			34.5	< .01

2.4.3. Duration effects

Figure 5 displays the duration of each syllable under different focus and modality conditions in the three sets of sentences. As can be seen, focus tends to lengthen not only the stressed syllable directly under focus, but also the following unstressed syllables of the focused word. Table 4 displays significant effects on the duration of stressed syllables in the key words under different focus, modality, and syllable position conditions based on two 3-way ANOVAs. The only significant effects are focus, syllable position, and their interactions. There is no modality effect on duration.

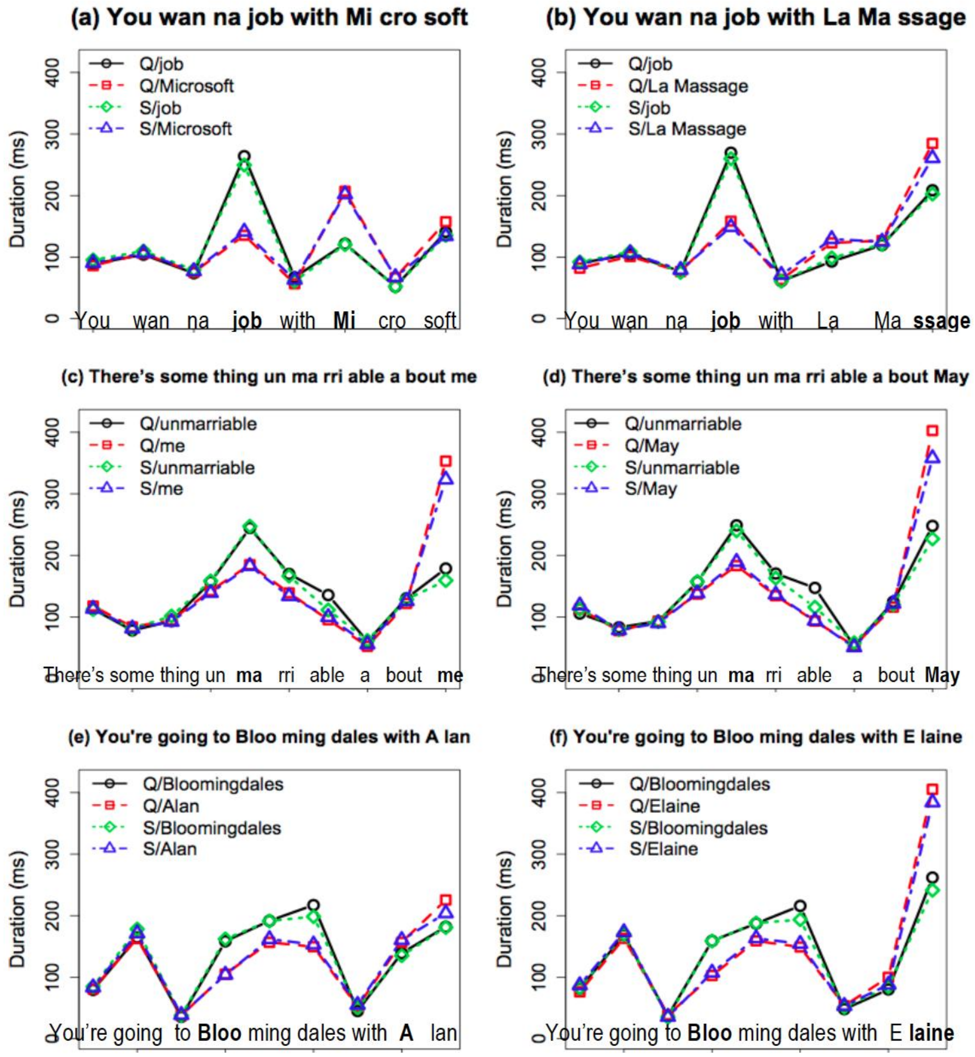


Figure 5. Durations of the syllables under different focus and modality conditions in English. In the legend, "S" stands for statement and "Q" for yes/no question. "Q/job" means a question with focus on "job", and so on.

Table 4. Significant effects of modality, focus, and syllable position on duration of the stressed syllables in sentence-medial and -final positions.

Significant effects	Sentence-medial		Sentence-final	
(*df* = 1, 4)	*F*	*p*	*F*	*p*
Focus	144.42	< .001	113.88	< .001
Syllable position	10.66	< .05	44.89	< .01
Focus * Syllable position	20.43	< .05	28.69	< .01

2.5. Discussion

The most striking results of this experiment are the strong and multiple interactions between focus, modality, and syllable position on all the acoustic measurements. Specifically, final-F_0 of post-focus syllables is lowered in statements, but raised in yes/no questions, which agrees with previous findings (14, 17), but is verified here with syllable-specific target values. In addition, final-velocity values demonstrate that F_0 in statements and yes/no questions in English is approximating different underlying pitch targets in stressed syllables of content words, with the former approaching a high-level or falling target (depending on the word's stress pattern, sentential position, and focus condition) and the latter approaching a rising target. This again agrees with the general observations in (17, 23), but is verified here with a measurement that will be shown later to be relevant for computational modeling. Furthermore, focus lengthens the focused word in both statements and yes/no questions, but no modality effect on duration is found in this experiment. This is also consistent with previous findings. (17)

In summary, Experiment 1 suggests the following conclusions about the prosody of General American English: 1) The F_0 difference between statements and yes/no questions becomes salient starting from the stressed syllable of the first content word, whether or not it is focused. This difference in modality is manifested in a shift of the underlying pitch targets of the stressed syllables from high-level or falling in statements to rising in yes/no questions. 2) Focus increases the pitch range of the focused word, lowers (in statements) or raises (in questions) that of the post-focus words, and leaves that of pre-focus words largely unaffected. 3) While focus tends to lengthen the durations of the on- and post-focus syllables in the focused word in both statements and yes/no questions, modality has no consistent effect on the duration of stressed syllables. 4) Intensity patterns in general follow those of F_0 variations due to both focus and modality. These patterns together clearly demonstrate evidence of *multi-componential coding* as well as *conditional allomorphs* of both focus and modality in English.

The strong interaction effects, which seem to have rendered the main effect of focus non-significant on both final-F_0 and final-velocity, together with the consistent focus effect on intensity, demonstrate why, when other crucial factors are not controlled, focus may show strong main effect on intensity, but not on F_0, as found in Kochanski et al. (2005). (36)

One of the most striking features of question intonation in English is the robust upshift of pitch range at the location of focus, as shown in Figure 2 and also in (17) and (14). When the word “job” is focused, for example, the average F_0 jump is 9.7 st within one syllable. It is possible that it is the weakening of the post-focus components that has allowed the high-pitched marking of interrogation to be fully realized. If that is the case, the neutral tone in Mandarin, which is also weak in articulation, (37) may also allow the same drastic post-focus pitch raising in questions. This is examined in Experiment 2.

3. Experiment 2 — Acoustic analysis of focus and question intonation in Mandarin

Similar to Experiment 1, the broad objective here is to examine for evidence of morpheme-like properties in both focus and question intonation in Mandarin, a tone language that differentiates words by both segmental and tonal contrasts. There are four full tones, High (H), Rising (R), Low (L) and Falling (F), and a neutral tone in Mandarin. The neutral tone has traditionally been considered as targetless, because its F_0 seems to be determined by the preceding tone. (38-39) Later experimental data suggest, however, that the neutral tone may have a mid target, and the contextual variability of its F_0 contour is due to its intrinsic weak articulatory strength. (37)

Although the basic features of Mandarin tones are largely preserved in connected speech, focus and modality have been found to modify the height and/or contour characteristics of the local tones, showing evidence of *multi-componential coding* of both functions. First, like in English, pitch range of the focused word in Mandarin is expanded, while that of post-focus words compressed, (25, 40) accompanied by similar changes in intensity. (19) Secondly, the overall F_0 values of the sentence-final full tones are higher in questions than in statements. (25, 41-47) Finally, F_0 contours of all the tones are tilted upward slightly at the end of a question. (25, 42-49)

Evidence of *conditional allomorphs* can be seen in the finding that at the statement-final position only a focused L tone is realized as falling-rising, while an unfocused L tone is falling. In contrast, the question-final L tone is always falling-rising regardless of whether it is focused. (25)

An initial comparison of Mandarin and English can already show some evidence of language specificity. As found in Liu and Xu (2005), even in a question, post-focus F_0 drops below the F_0 of the neutral-focus sentence, (25) which is different from the robust post-focus upshift in English questions as seen in Experiment 1. In that study, however, the sentences consisted of only full tones. It is possible that if neutral-tone syllables immediately follow a focused word, especially if the tone under focus is R, post-focus upshift of F_0 also occurs in Mandarin, because of the weak articulatory strength of the neutral tone. On the other hand, it is also possible that, because the neutral tone still has a target specification despite being weak, (37) its realization under the interaction of focus and interrogation is not fundamentally different from the full tones, except what can be predicted from its target as well as weak strength.

The aim of Experiment 2 is to add to previous findings by examining how the neutral tone in Mandarin interacts with focus and modality in terms of F_0 and duration, which has never been systematically examined. More specifically, we would like to find out if the two robust changes related to question intonation as seen in the English data also occur in Mandarin: a) post-focus upshift of pitch range, and b) rising shift of pitch target.[iii]

3.1. Materials

We composed two sets of sentences that ended with either neutral or H tones (Table 5). In each set of these sentences, the tones of the first and second syllables were always H and L, that of the third syllable varied across four full tones, those of the fourth, fifth, and sixth syllables were all neutral, and those of the seventh and eighth syllables were either both H or both neutral.

Table 5. Sentence structure of Experiment 2. The numbers at the end of the syllables represent the five tones: 0, 1, 2, 3, and 4 for N (Neutral), H (High), R (Rising), L (Low), and F (Falling), respectively.

Syllables 1-2	**Syllables 3-4**	**Syllables 5-8**
ta1 mai3 H L *“He bought”*	ma1 ma0 H N *“mothers’”*	men0 de0 le0 ma0 N N N N *“goodies”*
	ye2 ye0 R N *“grandpas’”*	
	nai3 nai0 L N *“grandmas’”*	men0 de0 mao1 mi1 N N H H *“kitten”*
	mei4 mei0 F N *“sisters’”*	

The following two conditions were imposed onto the above two sets of sentences so that a total of 32 combinations of the factors were created.

Focus: focus2 vs. focus3. The focus was either on the second syllable *mǎi*, which was separated from the neutral tones by a full tone, or on the third syllable *mā/yé/nǎi/mèi*, which was immediately adjacent to the neutral tones.

Modality: statement vs. question. Each utterance with the same components was produced with two alternate sentence types.

Each utterance was repeated five times by each subject, resulting in a total of 1280 sentences. The intended focus and modality were elicited by different prompt sentences (see Appendix 2 for details).

3.2. Subjects

Eight native speakers of Mandarin, 4 females and 4 males, served as subjects. They were either students at Yale University or residents in New Haven, Connecticut, who were born and raised in the city of Beijing where Mandarin is the vernacular. Aged between 23 and 34, they had no self-reported speech or hearing disorders.

3.3. Recording procedure and acoustic analysis

Recording was done in a sound-isolated booth at Haskins Laboratories, New Haven, Connecticut. Recording procedure was similar to that of Experiment 1, except that the utterances were directly digitized onto a computer hard disk. The utterances were digitized at 44.1 kHz sampling rate and 16-bit amplitude resolution, and were later re-sampled at 22.05 kHz. Analysis procedure was also similar to that in Experiment 1.

3.4. Results

3.4.1. Effects of focus and modality

Figure 6 displays time-normalized mean F_0 contours of the statements and questions with focus on either the second or third syllable of each sentence. Each contour was averaged logarithmically across 40 utterances (8 speakers × 5 repetitions). These plots were arranged in such a way as to highlight the effects of modality and tone of the final two syllables in the sentence. With regard to modality, questions clearly show higher F_0 than statements. However, there is an apparent lack of the kind of robust post-focus upshift in questions as in English. The greatest upshifts occur when the tone of syllable 3 is R or L. But this happens to both questions and statements, and F_0 eventually drops down after the delayed F_0 peaks rather than staying high in questions as in English. Also can be seen from Figure 6, starting from the focused item, the F_0 of questions becomes increasingly higher than that of statements, and the largest difference occurs at the end of a sentence.

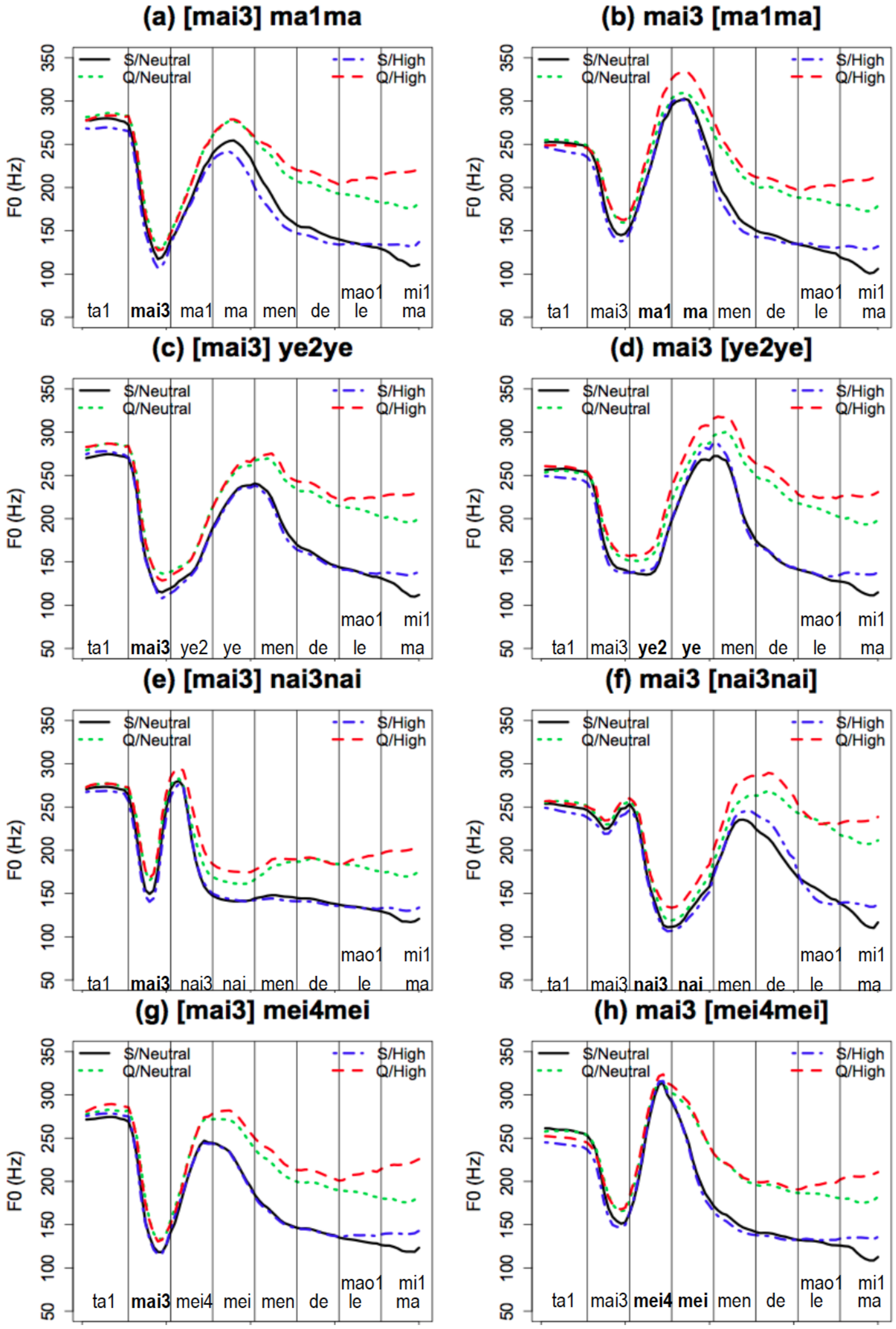

(a) [mai3] ma1ma
(b) mai3 [ma1ma]
(c) [mai3] ye2ye
(d) mai3 [ye2ye]
(e) [mai3] nai3nai
(f) mai3 [nai3nai]
(g) [mai3] mei4mei
(h) mai3 [mei4mei]
S/Neutral
Q/Neutral
S/High
Q/High
F0 (Hz)
50
100
150
200
250
300
350
ta1
mai3
ma1
ma
ye2
ye
nai3
nai
mei4
mei
men
de
mao1
le
mi1
ma

Figure 6. Time-normalized mean F_0 contours (averaged in logarithmic scale across 40 repetitions by 8 subjects) of Mandarin statements/questions with focus on "mǎi" (in (a), (c), (e), and (g)) and on "mā/yé/nǎi/mèi" (in (b), (d), (f), and (h)). In the legend, "S/Neutral" refers to a statement with a neutral-tone ending, "Q/High" a question with a HH-tone ending, and so on. The title "[mai3] ma1ma" in Figure 6(a) indicates that focus is on "mǎi" and the 3rd and 4th syllables of the sentence are "māma". The title "mai3 [ma1ma]" in Figure 6(b) indicates that focus is on "māma", the 3rd and 4th syllables of the sentence.

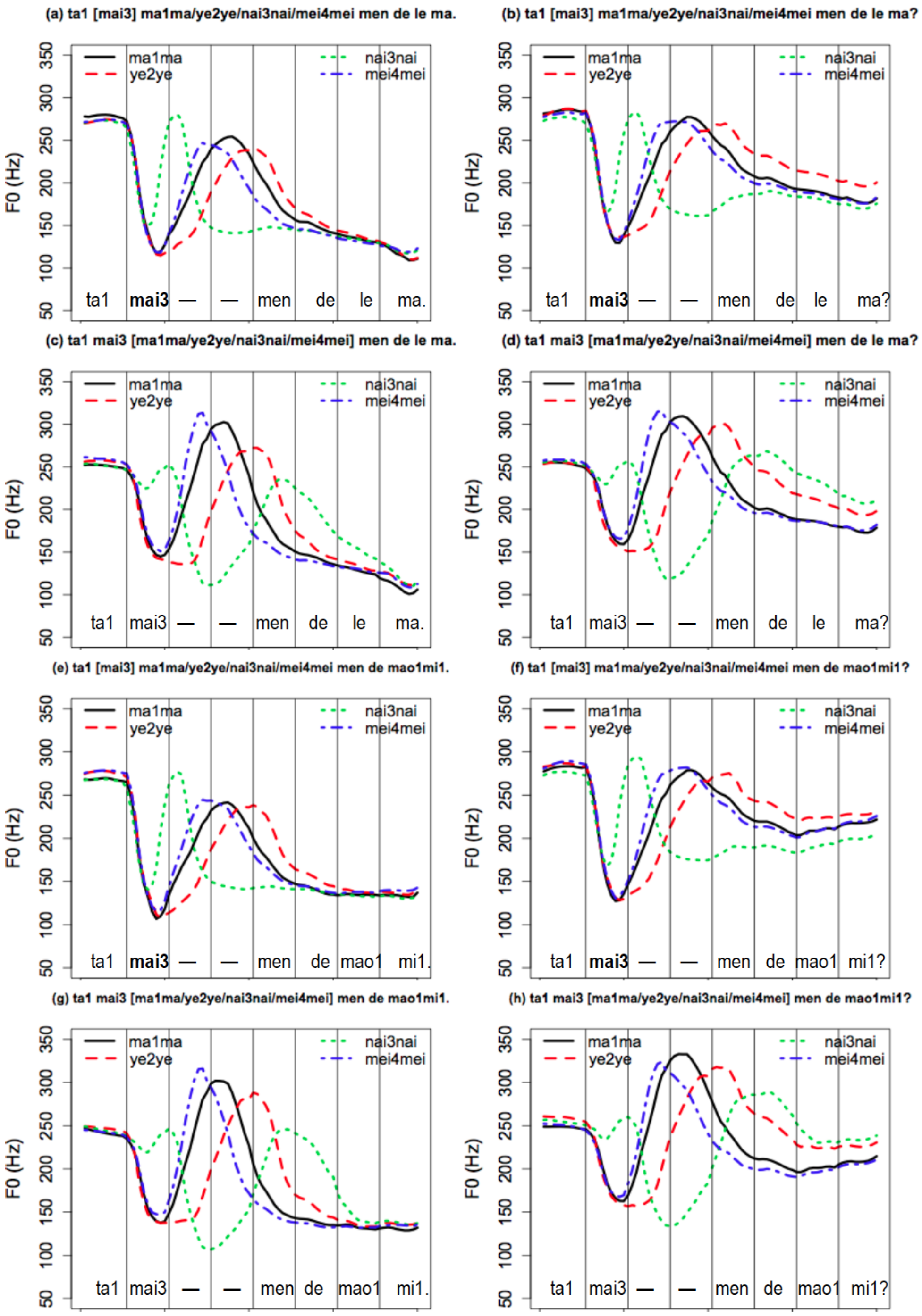
Table 6 shows results of repeated measures ANOVAs with *focus* (focus2/focus3), *modality* (statement/question), *full tone* (of syllable 3: H, R, L, or F), and *final tone* (H/neutral) on final-F_0 of syllables 3, 7, and 8. The statistics of only these three syllables are reported because they best reflect the effects of the main factors. As can be seen in the *modality* column, final-F_0 is significantly higher in questions than in statements in all three syllables. *Final tone* also has significant effects on all three syllables, but the effects are especially strong on the sentence-final syllable. This is partially reflected in the interaction between *modality* and *final tone* in syllable 7 (*le/māo*). There is no significant main effect of *focus*, but *focus* nevertheless has significant interactions with *full tone* and *modality*. In particular, the interaction of *focus* and *modality* for the last two syllables shows that post-focus F_0 is lowered less in questions than in statements.

Table 6. Significant effects from repeated measures ANOVAs on final-F_0 of syllable 3, 7 and 8 by focus (focus 2, focus 3), modality (S, Q), final tone (H, N), and tone of syllable 3 (H, R, L, F). Row 3 in each block shows the direction of the differences.

		Modality	Final tone	Full tone	Modality × Full tone	Modality × Final tone	Focus × Full tone	Focus × Modality
mā/ yé/ nǎi/ mèi	F	49.66	9.62	83.87	7.93		48.31	
	p	< .001	< .05	< .001	< .01		< .001	
		Q > S	H > N	(F, H) > (R, L)				
le/ māo	F	126.13	9.95	21.28	16.22	12.87	32.32	121.38
	p	< .001	< .05	< .001	< .001	< .05	< .001	< .001
		Q > S	H > N	(R, L) > (H, F)				
ma/ mī	F	126.41	28.92	8.25	5.37		29.38	20.09
	p	< .001	< .01	< .01	< .01		< .001	< .01
		Q > S	H > N	(R, L, F) > H				

3.4.2. Pitch target of the neutral tone

Figure 7 displays mean time-normalized F_0 contours again, but this time arranging them in such a way as to highlight the effects of the full tones of the third syllable on the subsequent neutral tones. The graphs are grouped according to whether the sentences are statements or questions, and whether the final two syllables have the neutral or H tone, as indicated at the top of each graph.



(a) ta1 [mai3] ma1ma/ye2ye/nai3nai/mei4mei men de le ma.

(b) ta1 [mai3] ma1ma/ye2ye/nai3nai/mei4mei men de le ma?

(c) ta1 mai3 [ma1ma/ye2ye/nai3nai/mei4mei] men de le ma.

(d) ta1 mai3 [ma1ma/ye2ye/nai3nai/mei4mei] men de le ma?

(e) ta1 [mai3] ma1ma/ye2ye/nai3nai/mei4mei men de mao1mi1.

(f) ta1 [mai3] ma1ma/ye2ye/nai3nai/mei4mei men de mao1mi1?

(g) ta1 mai3 [ma1ma/ye2ye/nai3nai/mei4mei] men de mao1mi1.

(h) ta1 mai3 [ma1ma/ye2ye/nai3nai/mei4mei] men de mao1mi1?

Figure 7. Time-normalized F_0 contours of Mandarin statements/questions alternating across "māma/yéye/nǎinai/mèimei" on the third and fourth syllables with focus on "mǎi" (in (a), (b), (e), and (f)) or on "mā/yé/nǎi/mèi" (in (c), (d), (g), and (h)). Figures 7(a-d) contain neutral-tone-ending sentences, and Figures 7(e-h) contain HH-tone-ending sentences.

As can be seen, despite the strong carryover influence from the preceding full tones, the F_0 contours of the neutral tone gradually converge over time and across the syllables, and the convergence is virtually complete by the end of the third neutral tone syllable in Figures 7a and 7e. Such convergence suggests that F_0 of the neutral tone is approaching a particular target rather than being fully determined by the preceding full tone. The slow rate of convergence indicates a weak articulatory strength associated with the neutral tone. (37) In Figure 7c and 7g, the convergence is not complete until the last neutral tone in Figure 7c, because after *nǎi* F_0 is actually rising, and the rise continues until the second or third neutral tone syllable. This rising F_0 contour after the L tone has been described as "post-low bouncing", (37) and has also been reported for Cantonese. (50) Figures 7a, c, e and g suggest that the effect is weak when the L tone itself is post-focus, but very strong when the L tone is focused. These variations, together with the rule-based tone change in the L tone of syllable 2 (L + L ➔ R + L) are captured by the significant interactions of focus × full tone for all three syllables. This is primarily due to the target shift of the second syllable and the "post-low bouncing" effect discussed above.

Figures 7b, d, f, and h show that the gradual convergence of F_0 across the neutral tone syllables is always incomplete when the sentences are questions. The deviant contours are those where the full tone is either L or R. This is reflected in the interaction of modality × full tone for all three syllables as shown in Table 6. These deviant contours seem to confirm the goal of our experimental design in eliciting an upward F_0 shift in the post-focus neutral tone. However, the amount of upshift elicited, as observed above, is still far from the drastic post-focus F_0 upshift in English questions.

The underlying pitch target of the neutral tone was further examined in terms of final-velocity. Repeated measures ANOVAs with gender, focus, modality, full tone, and final tone as independent variables indicate that final-velocity of the sentence-final syllable is only marginally affected by modality ($F(1,6) = 6.31$, $p = .05$). Thus, the pitch target of the neutral tone can potentially be inferred from the final-velocity value of the sentence-final neutral tone *ma*. Table 7 displays mean final velocities of the neutral tone *ma* and the H tone *mī*. In statements, the final velocities of the two tones are not significantly different from zero, or from each other. In questions, the final velocities of the two tones are again not significantly different from each other, although both are significantly greater than zero. Assuming that the pitch target of the H tone is static, which is uncontroversial, the pitch target of the neutral tone should also be static. The positive velocities in questions, though significantly higher than zero for both tones, are relatively small compared to the steep slopes of dynamic tones, (27) and could be attributed to the pitch range raising by the interrogative function which continues to increase toward the end of the sentence.

Table 7. Mean final velocities (st/s) of the sentence-final neutral tone and H tone in Mandarin statements and questions. The *t*-tests indicate whether these velocities are significantly different from zero or from each other.

Tone Intonation	N	H	N vs. H
Statement	-4.03 *t*(63) = -1.59 *p* > .05	1.16 *t*(63) = 1.22 *p* > .05	*t*(63) = 1.71 *p* > .05
Question	4.76 *t*(63) = 9.25 *p* < .001	3.89 *t*(63) = 10.98 *p* < .001	*t*(63) = -1.93 *p* > .05

As for the height of this static target, Chen and Xu (37) concluded that it should be mid because its height is half way between the maximum F_0 of the F tone and the minimum F_0 of the L tone in the same sentence position. In the current data, the mean F_0 of the neutral tone *ma* is significantly lower than that of the H tone *mī* (F(1,6) = 37.74, *p* < .001, 86.80 st < 89.05 st). However, unlike in Chen and Xu (2006), (37) there is no L tone at the end of the sentence in the current data to act as a reference to the floor of the pitch range. Thus no clear conclusion on this matter can be drawn here, although the current data are not incompatible with the idea of the neutral tone having a mid target.

Figure 8 shows the duration of each syllable under different focus and modality conditions in the neutral-tone- and H-tone-ending sentences. The significant effects on duration are shown in Table 8. As can be seen from the *focus* column, focused syllables are significantly longer than their non-focused counterparts, but focus has no significant effect on the duration of post-focus syllables. The *modality* column shows that the duration of the last syllable is significantly longer in questions than in statements, whereas that of the second to last shows the opposite pattern. For syllable 3, there are significant focus × modality, focus × full tone, and modality × full tone interactions. These are due to the fact that duration differences among the full tones are different across different focus and modality conditions. For syllable 7 *le/māo*, the significant modality × final tone interaction is because the duration of *māo* was longer in statements than in questions, while the duration of *le* did not differ between statements and questions.

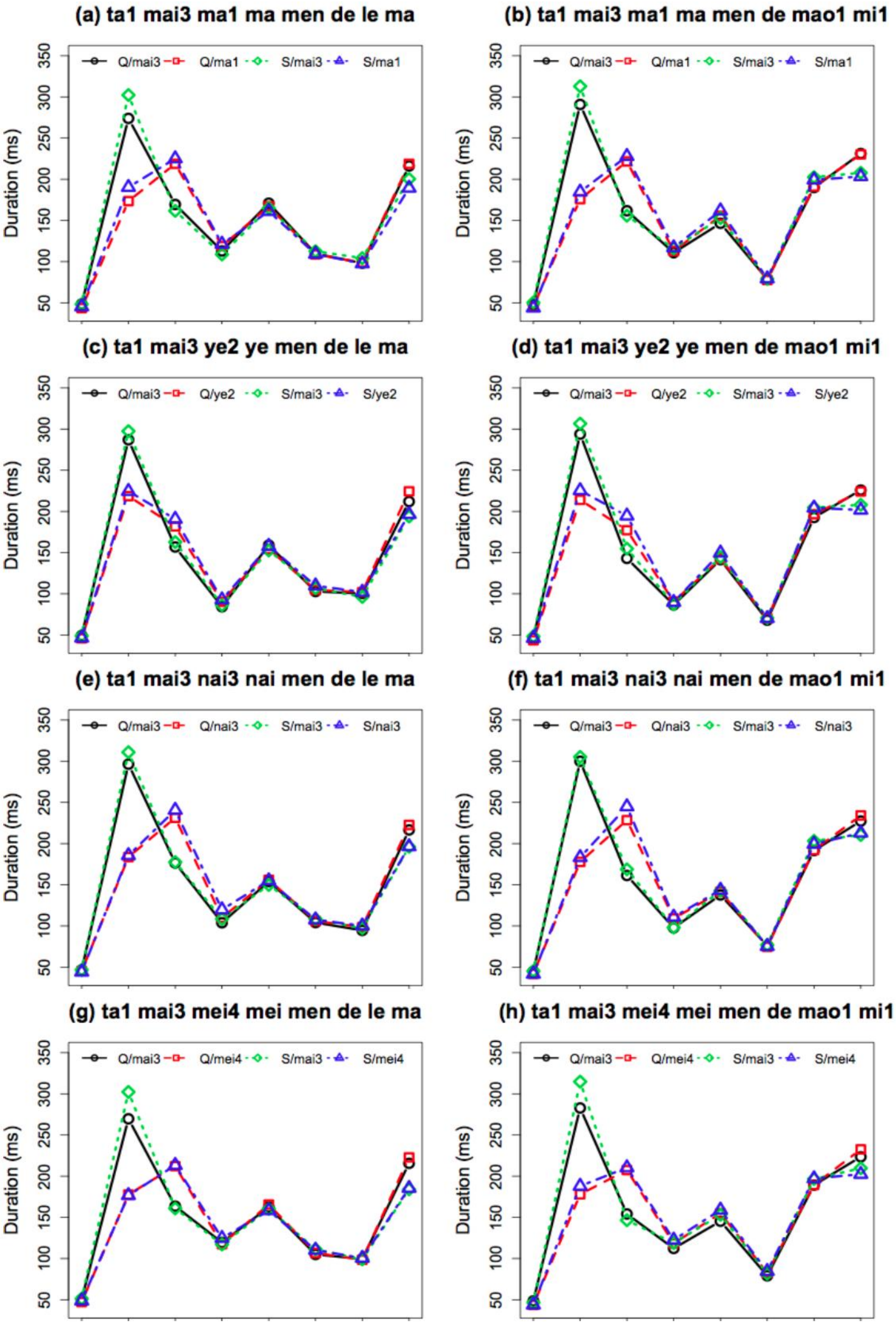

(a) ta1 mai3 ma1 ma men de le ma
(b) ta1 mai3 ma1 ma men de mao1 mi1
(c) ta1 mai3 ye2 ye men de le ma
(d) ta1 mai3 ye2 ye men de mao1 mi1
(e) ta1 mai3 nai3 nai men de le ma
(f) ta1 mai3 nai3 nai men de mao1 mi1
(g) ta1 mai3 mei4 mei men de le ma
(h) ta1 mai3 mei4 mei men de mao1 mi1
Duration (ms)
50
100
150
200
250
300
350
Q/mai3
Q/ma1
S/mai3
S/ma1
Q/ye2
S/ye2
Q/nai3
S/nai3
Q/mei4
S/mei4

Figure 8. Durations of the syllables under different focus and modality conditions in the neutral-tone-ending (in (a), (c), (e), and (g)) and HH-tone-ending sentences (in (b), (d), (f), and (h)). In the legend, "S" stands for statement and "Q" for question. "Q/mai3" means a question with focus on "mǎi", "S/ma1" means a statement with focus on "mā", and so on.

Table 8. Significant main effects and two-way interactions from repeated measures ANOVAs on duration of syllable 3, 7, and 8 by focus (Focus2: on *mǎi* vs. Focus3: on *mā/yé/nǎi/mèi*), modality (Question vs. Statement), final tone (H vs. N), and tone of syllable 3 (H, R, L, or F). Row 3 in each block shows the direction of the differences.

		Focus	Modality	Final tone	Tone	Modality × Tone	Modality × Final tone	Focus × Tone	Focus × Modality
mā/ yé/ nǎi/ mèi	F	33.95		24.32	7.47	8.08		12.90	14.43
	p	< .01		< .01	< .01	< .01		< .001	< .01
		A > N		H < N	L > (H,F) > R				
le/ māo	F		7.68	138.02			13.96		
	p		< .05	< .001			< .01		
			Q < S	H > N					
ma/ mī	F		51.09						
	p		< .001						
			Q > S						

3.5. Discussion

Like Experiment 1, this experiment has also found evidence of *multi-componential coding* of focus and modality in Mandarin. Also like English, the two functions interact with the lexical function, although in the case of Mandarin the critical lexical factor is tone. Question intonation in Mandarin exerts a nonlinear pitch increase starting from the focus position, as also found in Liu and Xu (2005), (25) and such increase is applied whether the sentence ends with a neutral or H tone. Unlike in English, however, post-focus pitch range is still lowered in questions in Mandarin, except that the amount of lowering is smaller than in a statement. This general pattern, which is consistent with the findings of Liu and Xu (25), did not change even when the post-focus syllables all had the neutral tone. In addition, focus lengthens the durations of the focused words in both statements and questions.

Traditionally considered to be "targetless", the neutral tone at the question-final position has been assumed to be fully free to carry the rising intonation and therefore should be high in pitch. (38, 51) However, the present data suggest that the neutral tone is not more effective than a full tone in manifesting question intonation, since the post-focus lowering as seen in the H-tone-ending questions also occurs in the neutral-tone-ending questions. This is probably because, although the F_0 trajectories of the neutral tone are heavily influenced by the preceding tone, they nevertheless gradually converge over time, indicating that the neutral tone is associated with an underlying pitch target, just like the other tones. According to the final-velocity values, the underlying target of the neutral tone is static and lower than the high target of the H tone. The very gradual approximation of the neutral tone target indicates that it is approached with a weak articulatory force, as suggested by Chen and Xu (2006). (37)

Overall, the above patterns indicate that, in Mandarin, focus and modality are independent and interactive intonational functions whose manifestations are achieved through modifying the pitch range of the local pitch target specified by the lexical tones, including the neutral tone. But such pitch range modifications do not seem to change the tonal pitch targets, even if the tone is weak like the neutral tone.

4. Experiment 3 — Quantitative modeling of English and Mandarin intonation

This experiment has four objectives. The first is to test whether *multi-componential coding* of focus and modality in English and Mandarin found by acoustic analyses in Experiments 1 and 2 can also be captured by computational modeling, using a newly developed modeling tool, PENTAtrainer, (52) based on the qTA model. (27) The second objective is to explore if extracted modeling parameters show resemblance to the target values in Experiments 1 and 2. The third objective is to use modeling to capture the functional interaction of focus, modality, lexical tone and lexical stress. Finally, the fourth objective is to explore the categoricalness of the prosodic functions by testing if the qTA parameters extracted from individual utterances can be prototyped into categorical targets, with which F_0 contours closely resembling those of natural speech can be predictively generated.

qTA is a mathematical implementation of the conceptual Target Approximation model. (28) It represents surface F_0 contours as the output of a third-order critically damped linear system in the form of

$$f_0(t) = x(t) + (c_1 + c_2 t + c_3 t^2)\, e^{-\lambda t}, \qquad (1)$$

The first term in parenthesis is the forced response, which is the pitch target, i.e., the desired underlying pitch trajectory associated with a syllable, which is specified not only in height (*b*, which is relative to sentence-initial F_0), but also in slope (*m*). The second term, consisting of the polynomial and the exponential, is the natural response, i.e., the transition from the previous articulatory state to the current pitch target. The model has only three free parameters, *m* and *b* which specify the pitch target, and λ which represents the rate or strength of target approximation. λ is inversely proportional to the time constant of the approximation process. The transient coefficients, c_1, c_2, and c_3, are jointly determined by the initial F_0 dynamic state of the syllable transferred from the preceding syllable and the pitch target. The initial dynamic state consists of F_0 level, $f_0(0)$, velocity, $f_0'(0)$, and acceleration, $f_0''(0)$, which are computed with the following formulae.

$$c_1 = f_0(0) - b \qquad (2)$$

$$c_2 = f'_0(0) + c_1\lambda - m \qquad (3)$$

$$c_3 = \left(f''_0(0) + 2c_2\lambda - c_1\lambda^2\right)/2 \qquad (4)$$

4.1. Method

The method applied here is similar to that in Prom-on et al. (27). The values of *m*, *b*, and λ were estimated through automatic analysis-by-synthesis. For each utterance in a corpus, the parameters were estimated syllable by syllable, starting from the beginning of the utterance. For each syllable, *m*, *b*, and λ are simultaneously estimated by searching for the value combination with the lowest sum of square error between the synthesized and original F_0 contours. Unlike in Prom-on et al. (27), no constraints were used to restrict the value of *m* in this experiment. During synthesis, the F_0 dynamic state (consisting of F_0, velocity, and acceleration) at each syllable offset is transferred to the next syllable as the initial condition, which is used in the parameter estimation for that syllable (equations 2-4).

This basic strategy had worked effectively for English and Mandarin sentences consisting of full tones. (27) For the neutral tone, however, in order to capture its weak strength, (25, 37, 53-54) the parameters need to be extracted by treating all the consecutive neutral tone syllables as a group during training. That is, for each utterance, all the neutral tone syllables are treated as having a common target. This way the weak strength of the neutral tone may be better captured according to our pilot data.

After obtaining the optimal target values from individual utterances, the parameters were then divided into categorical groups and for each group the median values of *m*, *b*, and λ were obtained. The categorical structure depends on the interaction between imposing communicative functions, e.g., stress, focus, and modality for English. To see the effect of individual difference, the parameters were also grouped in either speaker-dependent or speaker-independent fashion. Finally, the F_0 contour of each individual sentence was predictively synthesized using the categorical parameters based on the conditional combination of each of the syllables. As assessments of the modeling performance, sentence-level root-mean-square error (RMSE) and Pearson's correlation values were computed for each and every individual sentence.

4.2. Modeling results for English

Figure 9 displays the learned values of *b* and *m* in bar graphs, so as to make them comparable to the acoustic analysis shown in Figure 3. The values of post-focus *b* are highly positive in questions but highly negative in statements, which is consistent with the results of acoustic analysis shown in Figure 3a. Note that *b* is relative to sentence-initial F_0, so a negative value means that it is lower than the onset F_0 of the sentence. The values of *m* show a very similar pattern as the final-velocity values measured in Experiment 1, as shown in Figure 3b. In particular, stressed syllables show negative slopes in statements but positive slopes in questions, especially when they are word- or sentence-final.

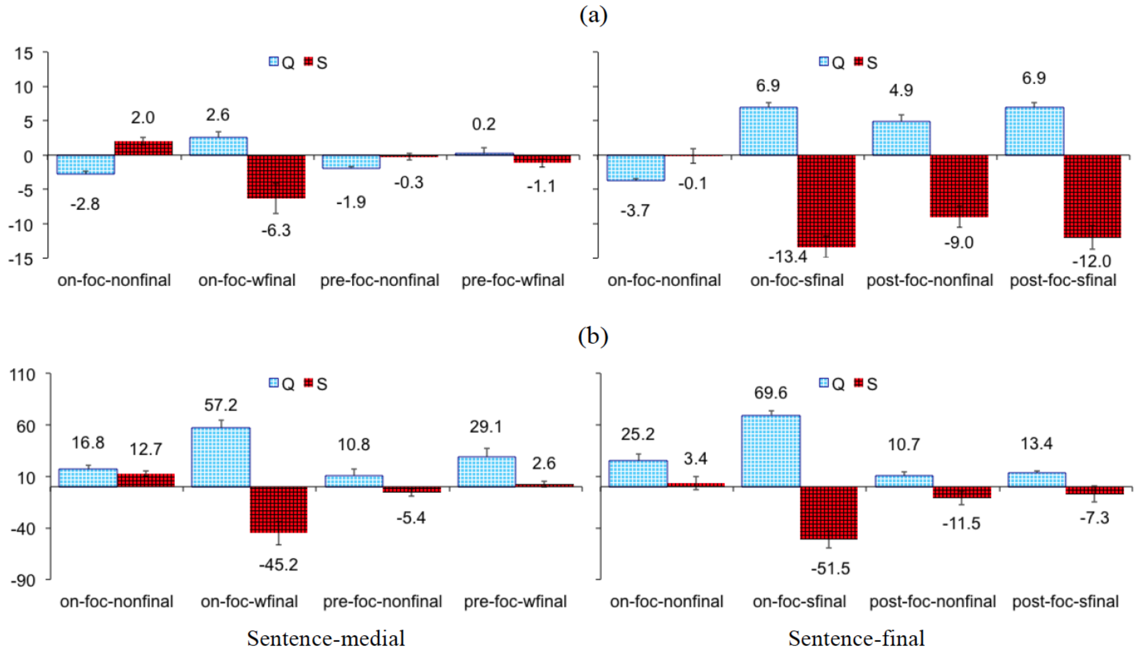


Figure 9. Values of *b* (top) and *m* (bottom) obtained through analysis-by-modeling on the English corpus.

Table 9 shows mean synthesis accuracies in terms of RMSE and correlation averaged according to synthesis conditions (rows) and whether the synthesis was speaker-dependent (left columns) or speaker-independent (right columns). The top left row shows the baseline speaker-dependent performance, comparing between the natural utterances and their conditional means, i.e., F_0 contours averaged across 5 repetitions of the same sentence in the same condition. The non-zero RMSE and imperfect correlation in this row indicate that random variations occurred across repetitions of the same sentence by the same speaker. The top right row shows baseline speaker-independent performance comparing between individual utterances and the cross-speaker conditional means. The accuracy is lower than the within speaker comparisons in the top left row. The second row shows very similar values to the top row, indicating that qTA parameters optimized for each original utterance are capable of accurately resynthesizing original F_0 contours. However, the number of unique parameters in this synthesis is very large. The third row shows that synthesis accuracy is lowered when using target parameters averaged across 8 repetitions of each identical sentence in the same condition, with increased RMSE and decreased correlation. When targets are grouped solely by lexical stress, as shown in row 4, the number of parameters is reduced to 70, but the accuracies are also much reduced. With the additions of focus and modality, synthesis accuracy increased steadily. The number of parameters also increased, but even with the most detailed grouping (final row), the parameter number is still much smaller than that of utterance-specific resynthesis or sentence-specific synthesis.

Table 9. Mean RMSE and correlation and their standard error values in synthesizing the English corpus. The first data row shows results of comparing individual sentence contours to mean sentence contours averaged across 8 repetitions by the same speaker (speaker dependent) or by all speakers (speaker independent).

Imposed Function	**Speaker Dependent**			**Speaker Independent**		
	Number of parameters	**RMSE (st)**	**Correlation**	**Number of parameters**	**RMSE (st)**	**Correlation**
Average vs. individual original	—	1.68	0.93	—	2.63	0.86
Resynthesis	8760	0.64 ± 0.05	0.975 ± 0.006	—	—	—
Resyntehsis with mean parameters	1080	1.90 ± 0.17	0.907 ± 0.015	216	2.64 ± 0.09	0.863 ± 0.011
Stress	70	3.97 ± 0.29	0.478 ± 0.028	14	3.98 ± 0.36	0.451 ± 0.023
Stress + Focus	135	3.99 ± 0.27	0.482 ± 0.028	27	4.01 ± 0.36	0.454 ± 0.024
Stress + Sentence	140	3.22 ± 0.19	0.750 ± 0.012	28	3.55 ± 0.17	0.731 ± 0.012
Stress + Focus + Sentence	270	2.56 ± 0.15	0.841 ± 0.014	54	3.00 ± 0.05	0.821 ± 0.013

4.3. Modeling results for Mandarin

Table 10 displays mean target parameters learned from the Mandarin data through qTA modeling. In the first data column, the learned values of *m* are positive, especially in the on-focus condition, which is not highly characteristic of H as a static tone. The reason is that the on-focus H tone only occurs in syllable 2, which is always preceded by the L tone. As a result, its F_0 contour is always rising despite the underlying static target, thus giving the training process little chance to discover the static target. For the R and F tones, *m* shows appropriate positive and negative values, reflecting their dynamic targets.

For *b*, the most notable is that in all the tones, the post-focus value is negative even in questions, which contrast sharply with the positive values in English questions shown in Figure 9, confirming the language difference shown in Experiments 1 and 2. Compared to H and L, the neutral tone shows intermediate values of *b* under similar conditions. In regard to λ, the most notable is that it is much smaller in the neutral tone than in the other tones, which demonstrates the effectiveness of the modeling strategy in capturing the weaker articulatory strength of the neutral tone.

Table 10. Mean target parameters of the Mandarin data learned through qTA modeling.

		m		*b*		λ	
Tone	Focus	Q	S	Q	S	Q	S
H	on	42	49	2.9	1.5	74	86
	post	23.5	25	-7.85	-3.25	60	63
	pre	-32	-32	0.0	0.2	69	72
R	on	86	78.5	-5	-3.7	28	37
	post	67	65	-7.4	-8.8	77	75
	pre	36	34	-0.7	-0.4	48	48
L	on	-16	-14	-13.7	-15.3	34	27
	post	-69	-79	-6.7	-9.9	67	54
	pre	0	-13	-10.9	-9.0	38	46
F	on	-10	-10	4.0	4.1	54	46
	post	32	27	-0.4	-1.8	78	85
Neutral	post	-2.5	0	-13.55	-5.9	12	12

Table 11 shows mean synthesis accuracies in a similar way as Table 9 for the English corpus. Like in Table 9, the accuracies of resynthesis are similar to the baseline comparisons in the top row, indicating that qTA parameters are capable of accurately regenerating all the details of the original F_0 contours. The third row shows that synthesis accuracy is lowered when using parameters averaged across 5 repetitions of each identical sentence in an identical condition. When targets are grouped solely by lexical tone, as shown in row 4, the number of parameters is reduced to 80, but the accuracies are also much reduced. With the additions of focus and modality, synthesis accuracy increased steadily, eventually reaching the same level of performance as resynthesis with mean parameters (row 3).

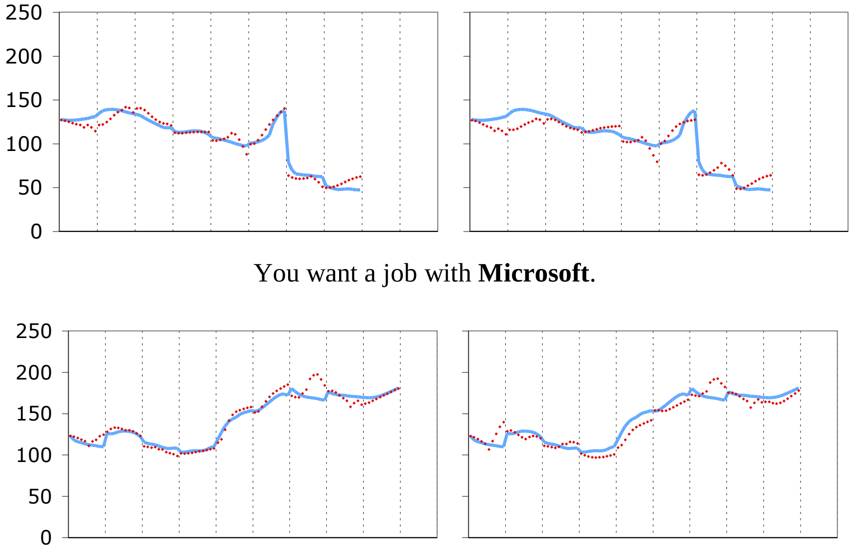
Table 11. Mean RMSE and correlation and their standard error values in synthesizing the Mandarin corpus. The first data row shows results of comparing individual sentence contours to mean sentence contours averaged across 5 repetitions by the same speaker (speaker dependent) or by all speakers (speaker independent).

Imposed Function	Speaker Dependent			Speaker Independent		
	Number of parameters	RMSE (st)	Correlation	Number of parameters	RMSE (st)	Correlation
Average vs. individual original	—	1.47 ± 0.18	0.97 ± 0.003	—	5.30 ± 0.88	0.92 ± 0.007
Resynthesis	10240	0.84 ± 0.07	0.946 ± 0.003	—	—	—
Resyntehsis with mean parameters	2048	3.36 ± 0.33	0.830 ± 0.012	256	3.64 ± 0.21	0.818 ± 0.011
Tone	80	4.34 ± 0.29	0.706 ± 0.022	10	4.46 ± 0.28	0.706 ± 0.016

Tone+Focus	136	3.90 ± 0.26	0.785 ± 0.016	17	3.99 ± 0.28	0.773 ± 0.020
Tone+Modality	160	3.71 ± 0.28	0.755 ± 0.025	20	4.00 ± 0.26	0.740 ± 0.018
Tone+Focus+Modality	272	3.26 ± 0.22	0.826 ± 0.017	34	3.66 ± 0.24	0.814 ± 0.016

4.4. Discussion

The results of this experiment show successful achievement of the experimental objectives. The modeling strategy based on multi-componential coding has led to successful modeling of focus and modality in both English and Mandarin; the automatically extracted model parameters show resemblance to the target values obtained in Experiments 1 and 2 by acoustic analysis; the importance of capturing the interactions between focus, modality, lexical tone and lexical stress is clearly demonstrated by Tables 9 and 11; and the categoricalness of focus and modality is demonstrated by the successful prediction of natural F_0 with prototyped categorical targets. Figure 10 shows sample English F_0 contours synthesized with the categorical parameters shown in Figure 9, together with the mean natural contours by one of the male speakers for whom the synthesis RMSE and correlation are in the middle range of all speakers. All contours are time-normalized averages across 8 repetitions, just like those of Figure 2a and 2d. The left two graphs are F_0 contours synthesized with speaker-specific parameters (i.e., those obtained from this speaker's data only), which show slightly tighter fit than in the right two graphs, where the contours are synthesized with parameters averaged across all five subjects. The overall close fit in both types of synthesis demonstrates that these categorical parameters have sufficiently captured how speakers control their production of intonation in this corpus. In addition, the modeling performance demonstrates that with an articulatorily-based model like qTA which requires only specifications of underlying targets in terms of height and slope, there is no need to explicitly process alignment of peaks and elbows. (55-59)



You want a job with **Microsoft**.

There is something **unmarriable** about May?

Figure 10. Sample synthetic F_0 contours compared to the natural contours. The solid curves are time-normalized natural contours averaged across 8 repetitions by one of the male subjects. The dotted curves are time-normalized synthetic curves using either speaker-specific (left) or cross-speaker categorical parameters.

With the modeling strategy that directly implements the morpheme-like properties found in the production experiments, continuous surface F_0 contours were accurately predicted. The contributions of each function can be specifically assessed, as shown in Tables 9 and 11. The best accuracy was achieved when all of the involved functions were implemented. Despite the reduction of the parameters to only a small set (which is smaller than any previous modeling studies we know, except Prom-on et al. (2009) (27)), the accuracy was almost equivalent to the non-categorized conditions (top 3 rows of Tables 9 and 11), indicating that the categorical parameters had captured the key properties of morpheme-like prosodic functions.

5. Further evidence

Beyond the results of the three experiments in the present study, additional evidence for the morpheme-like properties of prosodic functions can be found in many earlier findings as well as theoretical proposals, although those findings and proposals have not previously been interpreted this way. The following three sections will examine the earlier evidence from the perspective of the current paper.

5.1. Non-autonomy of components of focus and modality

Non-autonomy of components is in fact the single most important property of a morpheme, as it is directly derived from the definition that a morpheme is the smallest unit that carries meaning. For example, although “un-”, “certain” and “-ty” can each be isolated from “uncertainty” while still carrying meaning, “cer-’ and “-tain” are both meaningless once separated from each other. Evidence of this property is clearly seen in the case of focus. The perception study by Rump and Collier (60) has shown that a single early focus is marked by not only a medium to large F_0 peak on the focused word, but also an absence of any subsequence F_0 peak greater than 0.5 erb (≈ 0.71 st), as any later peak larger than that threshold would shift the perception to neutral or double focus (for replication in other languages, see (61)). This is consistent with the present as well as other findings mentioned earlier that in both English and Mandarin, on-focus expansion is consistently accompanied by post-focus compression. In other words, for the purpose of marking focus, an early “pitch accent” is not autonomous from the *absence* of later “pitch accents”.

Xu, Xu and Sun (2004) further showed that when a non-final word under focus was replaced by noise, Mandarin subjects could still judge the location of focus based on the intact post-focus words. Furthermore, when asked (without being told how) to imitate a sentence in which either the on-focus or post-focus word was replaced by noise, subjects nevertheless reproduced the *inaudible* on-focus expansion or post-focus compression. (62) These findings again show that to the naïve Mandarin speakers, on-focus expansion and post-focus compression are both intrinsic properties of focus that always co-occur, and that the total combination has developed into a stylized cohesive structure.

The stylized focus marking is also evident from findings that indicate target-like pitch range control. Rump and Collier (60) showed that both when trying to adjust the peak height of each of the key words while the height of the other was fixed, and when making passive perceptual judgments, Dutch listeners behaved as if they had in mind a categorical peak height specific to single initial focus, single final focus, neutral focus or double focus. The study further showed that this was in sharp contrast to the findings of the previous prominence experiments in which a scalar relationship was found when listeners were asked to pay attention to relatively small details of the speech signal. Chen and Gussenhoven (2008) further showed that when subjects had to put more emphasis on a word, the on-focus pitch height did not increase further beyond the level achieved in their first attempt. (63)

Contrary to the widely held assumption that words after the nuclear accent are fully deaccented, Pierrehumbert (1980:223) described what she referred to as echo accents: "Accentable syllables past the nuclear accent often carry a miniature replica of the nuclear accent. That is, in H* L- contours, one may see small peaks on accentable syllables following the H* nuclear accent; in L* H- contours, one [may] see small dips." (64) These observations are consistent with the finding that F_0 undulations related to lexical stress still occur in post-focus words (16) and the present finding that the type of pitch targets in pre-focus, on-focus and post-focus words in English all co-vary with modality (Figure 3). This is again evidence that statement and question, as two contrastive forms of modality, are both marked by stylized prosodic patterns consisting of multiple components, and that individual components are not autonomous from each other for the purpose of marking modality.

Furthermore, some proposed rules in the AM account of intonation can be reinterpreted as evidence of morpheme-like characteristics of focus and interrogation. The yes/no question intonation is transcribed as L* H-H%, where H- is a phrase accent that raises F_0 throughout post-focus region, and H% is a boundary tone. (8, 64) However, F_0 corresponding to H% is said to be further raised relative to H- by upstep — a phonetic implementation rule. (8, 65) What the present data show is that the H- phrase accent corresponds to the upshift of post-focus pitch range in English to mark a question. As such there is no issue of secondary association for accounting for the sometimes extensive plateau, (56, 66-67) because the temporal scope of post-focus region is determined directly by the location of focus in a sentence. The final rise corresponding to H% is part of the continuous F_0 increase to mark a question, which is in the opposite direction of statement marking by final lowering. (65) Taken together with the target shift corresponding to the echo accents, the raising of the entire post-focus pitch range is only part of the F_0 variations that serve to mark focus as well as question. In other words, none of these components can exist on their own once focus and modality are taken out of the picture.

5.2. Convergence of intonational meaning descriptions

The third line of additional evidence comes from previous theoretical proposals about morpheme-like meanings in prosody. As introduced at the beginning of this paper, according to Ladd (2008) the notion that "*the elements of intonation have morpheme-like meaning*" is widespread among linguists. (1) Those elements, however, refer to pitch accents, phrase accents and boundary tones, which are also viewed as phonological units not unlike segmental phonemes such as vowels and consonants. (1, 8, 64) The problem with this conceptualization is its incompatibility with the definition of morpheme, i.e., being the smallest meaningful units, and composed of variable number of phonological units like consonants, vowels, syllables and tones. Interestingly, however, a close look at the specific meanings proposed for those autonomous intonational units actually reveal that they describe the same kind of meanings that have been attributed to focus and modality, as shown with the following cases.

Case 1. Pierrehumber and Hirschberg (1990:289): "The H* accents … convey that the items made salient by the H* are to be treated as 'new' in the discourse. More generally, intonational phrases whose accents are all H* appear to signal to H [i.e., the hearer] that the open expression is to be instantiated by the accented items and the instantiated proposition realized by the phrase is to be added to H's mutual belief space." This description points to two pragmatic meanings: a) highlighting new information, and b) presenting it affirmatively. This is equivalent to marking a narrow focus in a statement. But as demonstrated by Experiment 1 as well as by previous empirical studies, (12, 14, 16), when elicited from speakers with the proper pragmatic contexts (e.g., prompt sentences), focus is associated not only with a prominent F_0 peak, i.e., the H* accent, but also with lengthened duration and increased intensity of the focused word, compressed and lowered pitch range of post-focus words, and a falling pitch target on the focused word-final stressed syllable (Figures 3 and 9). And it is such multi-componential coding that resembles the characteristic of lexical morphemes. As a further confirmation, AM theory also recognizes post-focus compression by referring to it as a L phrasal tone. According to Pierrehumbert and Hirschberg (1990:302), a "L phrasal tone emphasizes the separation of the current phrase from a subsequent phrase." From the functional perspective of the present paper, of course, the domain of post-focus compression, i.e., where the L phrase tone would be, is directly determined by the location of the focus relative to the entire sentence rather than by a phrase accent that is independent of focus.

Case 2. Pierrhumbert and Hirschberg (1990:291): "The L* accent marks items that S intends to be salient but not to form part of what S [i.e., the speaker] is predicating in the utterance. Schematically, one might say that S conveys that these items are not to be instantiated in the open expression that is to be added to H's mutual beliefs… In fact, S's motivation for marking these items as salient is the desire that H make such a predication." This complex description again points to two pragmatic meanings: a) highlighting a piece of information, and b) raising a doubt about it or hoping the listener to confirm it. In other words, it is equivalent to marking a narrow focus in a question. But as demonstrated by Experiment 1 as well as by previous empirical studies, (14, 17, present data) focus in a question is associated not only with a low rising F_0 contour, which is identified in AM theory as the L* accent, but also with lengthened duration of the focused word, compressed but raised pitch range of post-focus words, and assigning all stressed syllables a dynamic rising target (Figures 3 and 9).

Case 3: Pierrehumbert and Hirschberg (1990:304): "Boundary tones may also be H or L but have scope over the entire intonational phrase … a H boundary tone indicates that S wishes H to interpret an utterance with particular attention to subsequent utterances. A L boundary tone does not convey such directionality." In the same paper they have also mentioned that the H and L boundary tones typically occur in statements and questions, respectively, which is consistent with the findings of the present study. However, as demonstrated by Experiment 1 and 3 as well as previous function-based studies, (14, 17) sentence-final F_0 is not the sole prosodic correlate of the statement-question contrast in English. Rather, pitch targets of stressed syllables as well as pitch range of on-focus and post-focus components also systematically vary with the statement-question contrast. Thus, once again, it is the communicative function, referred to as modality in this study, that seems to encompass all the prosodic variations associated with the statement-question contrast.

It could be argued that even if pitch accents, phrase accents and boundary tones are not autonomous, they could be analogous to bound morphemes. But as we have just shown, their meanings as described by Pierrehumbert and Hirschberg (1990) are *synonymous with* rather than *independent of* the meanings of focus and modality. Thus, while the meanings of bound morphemes like "un-" and "-ty" are independent of the meaning of "certain" in "uncertainty", the meaning of post-nuclear L phrase accent is not independent of the meaning of the H* accent, as it is the combination of the two that carry the meaning of emphatic assertion, i.e., focus in statement.

In summary, the morpheme-like meanings previously proposed for the phonological intonational components do not seem to significantly differ from those associated with prosodic functions like focus and modality. But the multi-componential coding of the prosodic functions found in the present as well as previous empirical studies demonstrate that these functions are more analogous to lexical morphemes than the autonomous and largely phonetically defined intonational units. As a further indication, to our knowledge, there have been no experimental production studies that are able to specifically manipulate pitch accents, phrase accents or boundary tones except when training or imitation is involved. (56) All the existing experimental production studies, including those with the goal to specifically examine the phonological pitch units, (58) manipulated focus or modality or both.

5.3. Non-universality of focus and modality marking

Further evidence for the morpheme-like characteristics of at least some of the prosodic function can be seen in their variation across languages. By this we do not mean minor variations in terms of different ways in which focus and modality interact with each other, as seen in the present data. Rather, there is evidence in recent research indicating that the critical components of both focus and question intonation can be robustly different across languages of the world, and that the variability may have historical sources. In regard to focus, as seen in the present as well as previously reported data, post-focus compression — the reduction of pitch range and intensity in all post-focus words — occurs in both English and Mandarin. But such post-focus compression is recently found to be absent in many other languages. The first, and perhaps the most surprising finding is that it is absent in Taiwanese (also known as Southern Min), a Chinese language spoken in Taiwan and Hokkien, and Taiwan Mandarin, which is a language very similar to Beijing Mandarin. (19) Subsequently, a number of other languages in China, some belonging to the Chinese family, Cantonese, (68) others known as minority languages, including Yi, Deang and Wa, (69) were also found to lack PFC. In addition, a number of languages have been reported to lack prosodic focus in general, including Yucatec Maya, (70) Wolof, (71) and Chichewa, Chitumbuka, Durban Zulu, Hausa, Buli, and Northern Sotho. (72) Xu (2011) proposed that it is possible that all the languages that show PFC are historically related, and that PFC may have originated in the hypothetical proto-Nostratic language dating back about the end of the last Ice Age, i.e., 15,000-12,000 BC, probably spoken along the Fertile Crescent. (73) Although highly conjectural, this hypothesis has not yet been rejected by findings from on-going focus research so far.

Another line of research has shown evidence that even the most typical characteristic of question intonation found in many languages, namely, question-final rising F_0, corresponding to the H% boundary tone, is also not universal. Rialland (2009:928) reported that a group of languages in central Africa exhibit a so-called “lax” prosody: “Its typical realizations include a falling pitch contour, a sentence-final low vowel, vowel lengthening, and a breathy utterance termination produced by the gradual opening of the glottis." (74) She argued that this prosodic feature is "quite different from the well-known high-pitched or rising question prosody, common in Indo-European languages and elsewhere in the world and often considered to be a (near-) universal." These languages, which are of the Niger-Congo phylum, include the Gur, Kwa and Kru families, and some languages of the Nilo-Saharan phylum, and the Chadic family of the Afro-Asiatic phylum. (74) Again, all these languages are genetically or at least geographically related, suggesting a possible common origin of the "lax" prosody.

Thus there is evidence that the commonly observed prosodic markings of focus and modality may both have specific historical origins. While much more research is undoubtedly needed, the evidence so far is consistent with the idea that both structures are stylized cohesive patterns similar to lexical morphemes.

6. Additional issues

Despite the similarities with lexical morphemes detailed so far, prosodic functions also differ from lexical morphemes in one critical aspect. That is, as shown numerically in Experiment 3 and Prom-on et al. (2009), prosodic functions are typically modification functions. (27) This is not only in the fact that they are suprasegmental, but also that they achieve their marking by modifying existing local pitch targets already used by lexical functions, including lexical tone, (40) lexical stress, (16) and lexical quantity. (75) This is probably also why prosodic configurations have been so difficult to recognize. Unless examined under highly systematic experimental control, their configurations do not readily stand out, especially as cohesive functional wholes.

Also related to this issue is the fact that not all the prosodic functions seem to be as categorical as focus and modality. Boundary marking, which is mainly done through domain-final duration adjustment, has been found to be highly gradient. (76-77) Topic marking may also be gradient. (78-79) The gradience may be even finer when it comes to marking emotions and attitudes (80) a rarely recognized prosodic dimension. As the level of gradience increases, their resemblance to lexical morphemes also seems to be reduced. On the other hand, it is possible that the morphemic encoding strategy, even in the case of lexical morphemes, is only an evolved tendency rather than a designed property, just like segments and features which are also likely to be evolved tendency. (81) But even in the case of boundary marking and emotional expression, the most common feature of morphemic encoding, namely, *multi-componential* *coding*, is present. Boundary marking involves domain-final lengthening, pausing and sentence-final pitch lowering, (76-77, and 82) and emotional expressions involve manipulation of F_0, vocal tract length, voice quality, intensity and duration. (80)

Note that the multi-componential coding discussed in this paper is different from the multiplicity of cues at the phonemic level. In the latter case, any minimal contrast, such as voicing, may present multiple cues relevant to perception, as demonstrated by Lisker (1986). But most of those cues are the consequences of a single articulatory gesture. In the former, in contrast, the multiple components are not part of a single articulation, but separate articulations that are combined together only by the meaning function to be coded.

Finally, the present paper is only an initial attempt to clarify the link between meaning and form in prosody by identifying morpheme-like properties. We have exhausted neither all possible prosodic functions nor all relevant prosodic cues, because our discussion has to be based on available empirical data. Though preliminary and tentative, this effort is not intended to be a purely theoretical exercise; it is also driven by practical motivations. To improve the prosody capabilities of speech technology, for example, would it be preferable for text-to-speech and speech recognition systems to process autonomous intonational units and then find separate ways to link them to meanings as suggested by the Linguist's theory of intonational meaning? Or would it be preferable to directly process entities like focus and modality that are already functionally defined? Likewise, in language teaching or speech pathology, would it be more effective to focus on problems with autonomous phonological units or meaningful communicative functions? These are all still open questions, and answering them requires both empirical data and theoretical clarification of the meaning-form relation in prosody.

7. Conclusions

We have argued in this paper that prosodic functions exhibit properties similar to segmental morphemes: a) *multi-componential coding*, b) *conditional allomorphs*, c) *non-autonomy of components*, and d) *language-specificity with possible diachronic sources*. The key evidence presented involves functions like focus and modality. Focus involves on-focus increase of pitch range, duration, intensity, and high-frequency energy as well as post-focus decrease of F_0 and intensity. Sentence modality is marked not only by sentence-final F_0, but also by changes in pre-final F_0 and sometimes in local pitch targets as well. The production experiments demonstrated that when using experimental paradigms that manipulate communicative functions like focus and modality, multiple prosodic markers for each function were simultaneously elicited. The modeling experiment showed that the articulatorily-based qTA model could be trained on both experimental corpora to derive categorical parameters with values consistent with acoustic measurements from the production experiments. The morpheme-like properties of prosodic functions can also find support from evidence from previous research, especially in terms of *non-autonomy of components* and *language-specificity*. Finally, this study, for the first time, has developed a paradigm for comparing categorical prosodic properties obtained from acoustic analysis with those obtained from computational modeling. This has demonstrated that computational modeling is capable of serving as a tool for cross-validating empirical findings, and to do so at a level rarely seen before, namely, testing their relevance in predicting finely detailed surface prosody in natural speech, instead of merely validating specific hypotheses.

8. ACKNOWLEGEMENTS

This work was supported in part by the Economic and Social Research Council [Grant number PTA-026-27-2480-A to FL]; National Institutes of Health [Grant number R01DC03902 to YX] and Royal Society and the Royal Academy of Engineering [Newton International Fellowship to SP]. We also thank Professor Patrick Suppes (Center for the Study of Language and Information, Stanford University) for his financial support to FL. SP is supported by the Royal Society and the Royal Academy of Engineering through the Newton International Fellowship Scheme. Part of the results of Experiment 1 was presented at Interspeech 2007, Antwerp, Belgium. (53) Part of the results of Experiment 2 was presented at the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, (84) and part of the results of Experiment 3 was presented at the 17th International Congress of Phonetic Sciences, Hong Kong, China. (85) We owe our thanks to Gina-Anne Levow and John Goldsmith for insightful comments on an earlier version of the manuscript.

Appendix 1. Materials in Experiment 1. Words in boldface are focused. Prompt sentences are in parentheses, which were also read aloud by subjects.

Statement	Medial focus	Stressed final	(Not an **internship**.) You want a **job** with La Massage. (It's not **fate**.) There is something **unmarriable** about May. (It's not **Sears**.) You're going to **Bloomingdales** with Elaine.
		Un-stressed final	(Not an **internship**.) You want a **job** with Microsoft. (It's not **fate**.) There is something **unmarriable** about me. (It's not **Sears**.) You're going to **Bloomingdales** with Alan.
	Final focus	Stressed final	(Not **Microsoft**.) You want a job with **La Massage**. (It's not **me**.) There is something unmarriable about **May**. (It's not **Alan**.) You're going to Bloomingdales with **Elaine**.
		Un-stressed final	(Not **La Massage**.) You want a job with **Microsoft.** (It's not **you**.) There is something unmarriable about **me**. (It's not **Elaine**.) You're going to Bloomingdales with **Alan**.
Question	Medial focus	Stressed final	(Not an **internship**?) You want a **job** with La Massage? (It's not **fate**?) There is something **unmarriable** about May? (It's not **Sears**?) You're going to **Bloomingdales** with Elaine?
		Un-stressed final	(Not an **internship**?) You want a **job** with Microsoft? (It's not **fate**?) There is something **unmarriable** about me? (It's not **Sears**?) You're going to **Bloomingdales** with Alan?
	Final focus	Stressed final	(Not **Microsoft**?) You want a job with **La Massage**? (It's not **me**?) There is something unmarriable about **May**? (It's not **Alan**?) You're going to Bloomingdales with **Elaine**?
		Un-stressed final	(Not **La Massage**?) You want a job with **Microsoft?** (It's not **you**?) There is something unmarriable about **me**? (It's not **Elaine**?) You're going to Bloomingdales with **Alan**?

Appendix 2. Materials in Experiment 2. Words in boldface are focused. Sentences in parentheses are prompt sentences.

Statement	Focus on focus2	HH final	(Lǎowáng bú **mài** māma/yéye/nǎinai/mèimei men de māomī?) (‘Didn’t Laowang **sell** mothers’/grandpas’/grandmas’/sisters’ kittens?’) （ 老王 不 **卖** 妈妈 / 爷爷 / 奶奶 / 妹妹 们 的 猫咪？ ） tā **mǎi** māma/yéye/nǎinai/mèimei men de māomī. ‘He **bought** mothers’/grandpas’/grandmas’/sisters’ kittens.’ 他 **买** 妈妈 / 爷爷 / 奶奶 / 妹妹 们 的 猫咪。
		N final	(Lǎowáng bú **mài** māma/yéye/nǎinai/mèimei men de dōngxi?) (‘Didn’t Laowang **sell** mothers’/grandpas’/grandmas’/sisters’ goodies?’) （ 老王 不 **卖** 妈妈 / 爷爷 / 奶奶 / 妹妹 们 的 东西？ ） tā **mǎi** māma/yéye/nǎinai/mèimei men de le ma. ‘He **bought** mothers’/grandpas’/grandmas’/sisters’ [goodies].’ 他 **买** 妈妈 / 爷爷 / 奶奶 / 妹妹 们 的 了 嘛。
	Focus on focus3	HH final	(Lǎowáng bù mǎi **jiějie** men de māomī?) (‘Didn’t Laowang buy **older sisters’** kittens?’) （ 老王 不 买 **姐姐** 们 的 猫咪？ ） tā mǎi **māma/yéye/nǎinai/mèimei** men de māomī. ‘He bought **mothers’/grandpas’/grandmas’/younger sisters’** kittens.’ 他 买 **妈妈 / 爷爷 / 奶奶 / 妹妹** 们 的 猫咪。
		N final	(Lǎowáng bù mǎi **jiějie** men de dōngxi?) (‘Didn’t Laowang buy **older sisters’** goodies?’) （ 老王 不 买 **姐姐** 们 的 东西？ ） tā mǎi **māma/yéye/nǎinai/mèimei** men de le ma. ‘He bought **mothers’/grandpas’/grandmas’/younger sisters’** [goodies].’ 他 买 **妈妈 / 爷爷 / 奶奶 / 妹妹** 们 的 了 嘛。
Question	Focus on focus2	HH final	(Lǎowáng bú **mài** māma/yéye/nǎinai/mèimei men de māomī.) (‘Laowang didn’t **sell** mothers’/grandpas’/grandmas’/sisters’ kittens.’) （ 老王 不 **卖** 妈妈 / 爷爷 / 奶奶 / 妹妹 们 的 猫咪。 ） tā **mǎi** māma/yéye/nǎinai/mèimei men de māomī? ‘Did he **buy** mothers’/grandpas’/grandmas’/sisters’ kittens?’ 他 **买** 妈妈 / 爷爷 / 奶奶 / 妹妹 们 的 猫咪？
		N final	(Lǎowáng bú **mài** māma/yéye/nǎinai/mèimei men de dōngxi.) (‘Laowang didn’t **sell** mothers’/grandpas’/grandmas’/sisters’ goodies.’) （ 老王 不 **卖** 妈妈 / 爷爷 / 奶奶 / 妹妹 们 的 东西。 ） tā **mǎi** māma/yéye/nǎinai/mèimei men de le ma? ‘Did he **buy** mothers’/grandpas’/grandmas’/sisters’ [goodies]?’ 他 **买** 妈妈 / 爷爷 / 奶奶 / 妹妹 们 的 了 吗？
	Focus on focus3	HH final	(Lǎowáng bù mǎi **jiějie** men de māomī.) (‘Laowang didn’t buy **older sisters’** kittens.’) （ 老王 不 买 **姐姐** 们 的 猫咪。 ） tā mǎi **māma/yéye/nǎinai/mèimei** men de māomī?

			‘Did he buy **mothers’/grandpas’/grandmas’/younger sisters’** kittens?’ 他 买 **妈妈 / 爷爷 / 奶奶 / 妹妹** 们 的 猫咪？
		N final	(Lǎowáng bù mǎi **jiějie** men de dōngxi.) (‘Laowang didn’t buy **older sisters’** goodies.’) （老王 不 买 **姐姐** 们 的 东西。） tā mǎi **māma/yéye/nǎinai/mèimei** men de le ma? ‘Did he buy **mothers’/grandpas’/grandmas’/younger sisters’** [goodies]?’ 他 买 **妈妈 / 爷爷 / 奶奶 / 妹妹** 们 的 了 吗？

References

1. Ladd DR. Intonational phonology, Cambridge: Cambridge University Press; 2008.

2. Bolinger D. Intonation and Its Uses -- Melody in Grammar and Discourse, Stanford, California: Stanford University Press; 1989.

3. Gussenhoven C. Intonation and interpretation: Phonetics and Phonology In: Proceedings of The 1st International Conference on Speech Prosody, Aix-en-Provence, France, 2002. 47-57.

4. Gussenhoven C. The Phonology of Tone and Intonation: Cambridge University Press; 2004.

5. Hirschberg J. Communication and prosody: Functional aspects of prosody. Speech Commun. 2002;36:31-43.

6. Hirschberg J. Pragmatics and intonation. In: LR Horn and GL Ward, editors. The Handbook of Pragmatics, Oxford: Blackwell, 2004. p. 515-37.

7. Liberman MY. The intonational system of English [Ph.D. Dissertation]. M.I.T.; 1975.

8. Pierrehumbert J and Hirschberg J. The meaning of intonational contours in the interpretation of discourse. In: PR Cohen, J Morgan and ME Pollack, editors. Intentions in Communication, Cambridge, Massachusetts: MIT Press, 1990. p. 271-311.

9. Steedman M. Information Structure and the Syntax-Phonology Interface. Ling. Inq. 2000;31:649-89.

10. Xu Y. Speech melody as articulatorily implemented communicative functions. Speech Commun. 2005;46:220-51.

11. Spencer A. Morphological theory: An introduction to word structure in generative grammar, Oxford: Blackwell; 1991.

12. Cooper WE, Eady SJ and Mueller PR. Acoustical aspects of contrastive stress in question-answer contexts. J. Acoust. Soc. Am. 1985;77:2142-56.

13. Breen M, Fedorenko E, Wagner M and Gibson E. Acoustic correlates of information structure. Language and Cognitive Processes 2010;25:1044-98.

14. Pell MD. Influence of emotion and focus on prosody in matched statements and questions. J. Acoust. Soc. Am. 2001;109:1668-80.

15. Heldner M. On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. J. Phonetics 2003;31:39–62.

16. Xu Y and Xu CX. Phonetic realization of focus in English declarative intonation. J. Phonetics 2005;33:159-97.

17. Eady SJ and Cooper WE. Speech intonation and focus location in matched statements and questions. J. Acoust. Soc. Am. 1986;80:402-16.

18. Rialland A. African “lax” question prosody: its realisations and its geographical distribution. Lingua 2009;119:928-49.

19. Xu Y, Chen S-w and Wang B. Prosodic focus with and without post-focus compression (PFC): A typological divide within the same language family? The Linguistic Review 2012;29:131-47.

20. Bolinger D. Intonation across languages. In: JH Greenberg, editors. Universals of human language, Phonology V. 2: Stanford University Press, 1978. p. 471-523.

21. Gårding E and Abramson AS. A study of the perception of some American English intonation contours. Studia Linguistica 1965;19:61-79.

22. O'Shaughnessy D. Linguistic features in fundamental frequency patterns. J. Phonetics 1979;7:119-45.

23. O'Shaughnessy D and Allen J. Linguistic modality effects on fundamental frequency in speech. J. Acoust. Soc. Am. 1983;74:1155-71.

24. Thorsen NG. A study of the perception of sentence intonation — Evidence from Danish. J. Acoust. Soc. Am. 1980;67:1014-30.

25. Liu F and Xu Y. Parallel encoding of focus and interrogative meaning in Mandarin intonation. Phonetica 2005;62:70-87.

26. Fry DB. Experiments in the perception of stress. Lang. Speech 1958;1:126-52.

27. Prom-on S, Xu Y and Thipakorn B. Modeling tone and intonation in Mandarin and English as a process of target approximation. J. Acoust. Soc. Am. 2009;125:405-24.

28. Xu Y and Wang QE. Pitch targets and their realization: Evidence from Mandarin Chinese. Speech Commun. 2001;33:319-37.

29. Boersma P. Praat, a system for doing phonetics by computer. Glot International 2001;5:9/10:341-5.

30. Xu Y. ProsodyPro.praat. Available from: <http://www.phon.ucl.ac.uk/home/yi/ProsodyPro/>. 2005-2013.

31. Laniran YO and Clements GN. Downstep and high raising: interacting factors in Yoruba tone production. J. Phonetics 2003;31:203-50.

32. Xu Y and Liu F. Tonal alignment, syllable structure and coarticulation: Toward an integrated model. Italian Journal of Linguistics 2006;18:125-59.

33. Xu Y and Liu F. Determining the temporal interval of segments with the help of F0 contours. J. Phonetics 2007;35:398-420.

34. Gauthier B, Shi R and Xu Y. Learning phonetic categories by tracking movements. Cognition 2007;103:80-106.

35. Stevens KN. Acoustic Phonetics, Cambridge, MA: The MIT Press; 1998.

36. Kochanski G, Grabe E, Coleman J and Rosner B. Loudness predicts prominence: Fundamental frequency lends little. J. Acoust. Soc. Am. 2005;118:1038-54.

37. Chen Y and Xu Y. Production of weak elements in speech -- Evidence from f0 patterns of neutral tone in standard Chinese. Phonetica 2006;63:47-75.

38. Chao YR. Tone and intonation in Chinese. Bulletin of the Institute of History and Philology 1933;4:121-34.

39. Chao YR. A Grammar of Spoken Chinese, Berkeley, CA: University of California Press; 1968.

40. Xu Y. Effects of tone and focus on the formation and alignment of F0 contours. J. Phonetics 1999;27:55-105.

41. Ho AT. Intonation variation in a Mandarin sentence for three expressions: interrogative, exclamatory and declarative. Phonetica 1977;34:446-57.

42. Lin M. On production and perception of boundary tone in Chinese intonation In: Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, 2004. 125-9.

43. Ni J-F and Kawai H. Pitch targets anchor Chinese tone and intonation patterns In: Proceedings of International Conference on Speech Prosody 2004, Nara, Japan, 2004. 95-8.

44. Rumjancev MK. Ton i intonacija v sovremennom kitajskom jazyke (Tone and Intonation in Modern Chinese), Moscow: Izdatel'stvo Moskovskogo Universiteta; 1972.

45. Shen XS. The Prosody of Mandarin Chinese, Berkeley: University of California Press; 1990.

46. Yuan J. Intonation in Mandarin Chinese: Acoustics, perception, and computational modeling [Unpublished dissertation]. Cornell University, Ithaca, NY.; 2004.

47. Zeng X-L, Martin P and Boulakia G. Tones and intonation in declarative and interrogative sentences in Mandarin In: Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, 2004. 235-8.

48. Ho AT. Mandarin tones in relation to sentence intonation and grammatical structure. Journal of Chinese Linguistics 1976;4:1-13.

49. Ho AT. The acoustic variation of Mandarin tones. Phonetica 1976;33:353-67.

50. Gu W and Lee T. Effects of Tone and Emphatic Focus on Speech Prosody – A Comparison between Standard Chinese and Cantonese In: Proceedings of PCC2008, Beijing, 2008.

51. Qi S. Hanyu de zidiao, tingdun yu yudiao de jiaohu guanxi [The interaction among lexical tones, pause, and intonation in Chinese]. 中国语文 1956;10:10-3.

52. Xu Y and Prom-on S. PENTAtrainer1.praat. Available from: <http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/>. 2010-2012.

53. Liu F and Xu Y. The neutral tone in question intonation in Mandarin In: Proceedings of Interspeech 2007, Antwerp, 2007. 630-3.

54. Liu F. Intonation systems of Mandarin and English: A functional approach [PhD dissertation]. The University of Chicago; 2009.

55. Arvaniti A, Ladd DR and Mennen I. Stability of tonal alignment: the case of Greek prenuclear accents. J. Phonetics 1998;36:3-25.

56. Barnes J, Veilleux N, Brugos A and Shattuck-Hufnagel S. Turning points, tonal targets, and the English L- phrase accent. Language and Cognitive Processes 2010;25:982-1023.

57. Ladd DR, Mennen I and Schepman A. Phonological conditioning of peak alignment in rising pitch accents in Dutch. J. Acoust. Soc. Am. 2000;107:2685-96.

58. Shue Y-L, Shattuck-Hufnagel S, Iseli M, Jun S-A, Veilleux N and Alwan A. On the acoustic correlates of high and low nuclear pitch accents in American English. Speech Commun. 2010;52:106-22.

59. Welby P. The role of early fundamental frequency rises and elbows in French word segmentation. Speech Commun. 2007;49:28-48.

60. Rump HH and Collier R. Focus conditions and the prominence of pitch-accented syllables. Lang. Speech 1996;39:1-17.

61. Mixdorff H. Quantitative tone and intonation modeling across languages In: Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, 2004. 137-42.

62. Xu Y, Xu CX and Sun X. On the Temporal Domain of Focus In: Proceedings of International Conference on Speech Prosody 2004, Nara, Japan, 2004. 81-4.

63. Chen Y and Gussenhoven C. Emphasis and tonal implementation in Standard Chinese. J. Phonetics 2008;36:724-46.

64. Pierrehumbert J. The Phonology and Phonetics of English Intonation [Ph.D. dissertation]. MIT, Cambridge, MA. [Published in 1987 by Indiana University Linguistics Club, Bloomington]; 1980.

65. Liberman M and Pierrehumbert J. Intonational invariance under changes in pitch range and length. In: M Aronoff and R Oehrle, editors. Language Sound Structure, Cambridge, Massachusetts: M.I.T. Press, 1984. p. 157-233.

66. Grice M, Ladd DR and Arvaniti A. On the place of phrase accents in intonational phonology. Phonology 2000;17:143-85.

67. Pierrehumbert J and Beckman M. Japanese Tone Structure, Cambridge, MA: The MIT Press; 1988.

68. Wu WL and Xu Y. Prosodic Focus in Hong Kong Cantonese without Post-focus Compression In: Proceedings of Speech Prosody 2010, Chicago, 2010.

69. Wang B, Wang L and Kadir T. Prosodic encoding of focus in six languages in China In: Proceedings of The 17th International Congress of Phonetic Sciences, Hong Kong, 2011. 144-7.

70. Kügler F and Skopeteas S. On the universality of prosodic reflexes of contrast: The case of Yucatec Maya In: Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, 2007.

71. Rialland A and Robert S. The intonational system of Wolof. Linguistics 2001;39:893–939.

72. Zerbian S, Genzel S and Kügler F. Experimental work on prosodically-marked information structure in selected African languages (Afroasiatic and Niger-Congo) In: Proceedings of Speech Prosody 2010, Chicago, 2010. 100976:1-4.

73. Xu Y. Post-focus compression: Cross-linguistic distribution and historical origin In: Proceedings of The 17th International Congress of Phonetic Sciences, Hong Kong, 2011. 152-5.

74. Rialland A. African "lax" question prosody: its realisations and its geographical distribution. Lingua 2009;119:928-49.

75. Suomi K. On the tonal and temporal domains of accent in Finnish. J. Phonetics 2007;35:40-55.

76. Wagner M. Prosody and Recursion [Ph.D. Dissertation]. Massachusetts Institute of Technology; 2005.

77. Xu Y and Wang M. Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. J. Phonetics 2009;37:502-20.

78. Wang B and Xu Y. Differential prosodic encoding of topic and focus at sentence initial position in Mandarin Chinese. J. Phonetics 2011;39:595-611.

79. Swerts M and Ostendorf M. Prosodic and lexical indications of discourse structure in human-machine interactions. Speech Commun. 1997;22:25-41.

80. Xu Y, Kelly A and Smillie C. Emotional expressions as communicative signals. In: S Hancil and D Hirst, editors. Prosody and Iconicity: Benjamins, 2013. pp. 33-60.

81. Blevins J. Duality of patterning: Absolute universal or statistical tendency? langcog 2012;4:275-96.

82. Byrd D, Krivokapić J and Lee S. How far, how long: On the temporal scope of phrase boundary effects. J. Acoust. Soc. Am. 2006;120:1589-99.

83. Lisker L. "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. Lang. Speech 1986;29:3-11.

84. Liu F and Xu Y. Question intonation as affected by word stress and focus in English In: Proceedings of The 16th International Congress of Phonetic Sciences, Saarbrücken, 2007. 1189-92.

85. Prom-on S, Liu F and Xu Y. Functional modeling of tone, focus and sentence type in mandarin Chinese In: Proceedings of The 17th International Congress of Phonetic Sciences, Hong Kong, 2011. 1638-41.

[i] Previous research has shown that it makes little difference whether the leading sentences are uttered by the subject (Xu, 1999), the experimenter (Xu et al., 2012) or are pre-recorded (Xu & Xu, 2005).

[ii] Overall ANOVAs were also done for the effect of gender. However, besides having significantly different final-F_0, male and female speakers do not differ significantly on duration for most of the syllables, and there was little interaction of gender with other factors.

[iii] Given what we already know about intensity variation as a function of focus (Xu et al., 2012), and in the interest of space, intensity was not analyzed in this experiment.