Modeling tone and intonation in Mandarin and English as a process of target approximation

Santitham Prom-on[a)]
Computer Engineering Department, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand

Yi Xu[b)]
Department of Speech, Hearing and Phonetic Sciences, University College London, London WC1N 2PF, United Kingdom

Bundit Thipakorn[c)]
Computer Engineering Department, King Mongkut's University of Technology Thonburi, Bangkok 10140, Thailand

(Received 14 September 2007; revised 4 October 2008; accepted 6 November 2008)

This paper reports the development of a quantitative target approximation (qTA) model for generating F_0 contours of speech. The qTA model simulates the production of tone and intonation as a process of syllable-synchronized sequential target approximation [Xu, Y. (2005). "Speech melody as articulatorily implemented communicative functions," Speech Commun. **46**, 220–251]. It adopts a set of biomechanical and linguistic assumptions about the mechanisms of speech production. The communicative functions directly modeled are lexical tone in Mandarin and lexical stress in English and focus in both languages. The qTA model is evaluated by extracting function-specific model parameters from natural speech via supervised learning (automatic analysis by synthesis) and comparing the F_0 contours generated with the extracted parameters to those of natural utterances through numerical evaluation and perceptual testing. The F_0 contours generated by the qTA model with the learned parameters were very close to the natural contours in terms of root mean square error, rate of human identification of tone, and focus and judgment of naturalness by human listeners. The results demonstrate that the qTA model is both an effective tool for research on tone and intonation and a potentially effective system for automatic synthesis of tone and intonation. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3037222]

PACS number(s): 43.70.Bk, 43.72.Ja, 43.70.Fq, 43.70.Kv [AE] Pages: 405–424

I. INTRODUCTION

Quantitative modeling is one of the most rigorous means of testing our understanding of a natural phenomenon. This is particularly true if the model is built directly on assumptions that closely reflect the contested view about the mechanisms underlying the phenomenon. Modeling can also help to improve our knowledge by forcing us to make our theoretical postulations as explicit as possible. Thus for improving our understanding of human speech, quantitative modeling is also indispensable. In the present paper we report the results of an attempt to simulate tone, stress, and focus in Mandarin and English with a quantitative model that generates surface F_0 contours through the process of target approximation (TA) (Xu and Wang, 2001). Our goal is not only to develop a robust quantitative model applicable in speech technology but also to test our understanding of tone and intonation accumulated in recent years (Xu and Wang, 2001; Xu, 2005).

There have been many attempts over the past decades to build a robust model capable of simulating various prosodic phenomena through F_0 modeling (Bailly and Holm, 2005; Fujisaki, 1983; Fujisaki *et al.*, 2005; Hirst and Espesser, 1993; Kochanski and Shih, 2003; Ni *et al.*, 2006; Pierrehumbert, 1981; Taylor, 2000; van Santen and Möbius, 2000). These approaches can be divided into two general categories, namely, those that model F_0 contours directly and those that attempt to simulate the underlying mechanisms of F_0 production. Models belonging to the first category are derived mainly based on the shape of the F_0 contours, with minimal consideration about the articulatory process of F_0 production. These include the quadratic spline model (Hirst and Espesser, 1993), the Pierrehumbert model (Pierrehumbert, 1981), the tilt model (Taylor, 2000), the linear alignment model (van Santen and Möbius, 2000), the superposition of functional contours (SFC) model (Bailly and Holm, 2005), and the tone transformation model (Ni *et al.*, 2006). The quadratic spline model interpolates peaks and valleys of F_0 contours with a quadratic spline function while the Pierrehumbert model interpolates F_0 between adjacent peaks and valleys using a linear or sagging function. The tilt model generates F_0 from the tilt parameters which describe the shapes of F_0 in each intonational event, e.g., pitch accent and boundary tone. The F_0 contour of an utterance is represented by a series of these intonational events. The linear alignment model uses curve classes as templates, warping and then combining these

[a)]Electronic mail: santitham@cpe.kmutt.ac.th
[b)]Electronic mail: yi.xu@ucl.ac.uk
[c)]Electronic mail: bundit@cpe.kmutt.ac.th

curve classes superpositionally to generate F_0 contours. The SFC model simulates intonation by superpositionally combining multiple elementary contours that are functionally defined. The tone transformation model generates F_0 by specifying tone-related turning points and connecting them with a truncated second-order response function, and then imposing a global function onto the local tonal shapes. Although models in this category can represent F_0 contours at a certain level of accuracy, they do not separate surface patterns that carry intended information from those that are due to articulatory mechanisms. As a result, they have to either ignore most of the microvariations due to articulation, as done in various stylization strategies (d'Alessandro and Mertens, 1995; 't Hart *et al.*, 1990), or simulate all surface F_0 patterns directly as just described.

Models belonging to the second category are based on assumptions about the process of F_0 production. Examples in this category are the soft-template model (Kochanski and Shih, 2003; Kochanski *et al.*, 2003) and the command-response (CR) model (Fujisaki, 1983; Fujisaki *et al.*, 2005). The soft-template markup language (Stem-ML), based on a soft-template model, describes F_0 contours as resulting from realizing underlying tonal templates with different amounts of muscle forces under the physical constraint of smoothness (Kochanski *et al.*, 2003). The smoothness constraint guarantees continuous connections between adjacent templates, and the varying muscle force determines the degree to which the shape of each template is preserved in the surface F_0 under the influence of neighboring tones that are either adjacent or far away, and either preceding or following the targeted template. Stem-ML uses the optimization modeling approach for F_0 realization which requires sophisticated and complex error minimization. Even though the assumptions of Stem-ML are motivated by physical mechanisms, it requires complex mathematical translation from articulatory constraints into effort and error constraints of optimization modeling. The CR model (Fujisaki, 1983; Fujisaki *et al.*, 2005) offers, in our view, the most plausible physiological and physical simulations of the tension control mechanisms of the vocal folds compared to other models, and thus formulates the closest approximation of natural F_0 contours to date (Fujisaki *et al.*, 2005; Gu *et al.*, 2007), although certain aspects of it are still not satisfactory. The model represents surface F_0 as the logarithmic sum of phrase components and accent or tone components. The phrase components, which are assumed to be produced by the contraction of the pars obliqua of the cricothyroid (CT) muscle, represent the global contours of the utterance and are generated by a sequence of impulse response functions. The accent components, which are assumed to be produced by the contraction of the pars recta of the CT muscle, represent the local contours of the utterance and are generated by a sequence of step response functions. The CR model is thus based on the assumption that individual muscles are controlled separately. This assumption, however, is inconsistent with the finding that muscles are controlled as functional groups rather than individually (Bernstein, 1967; Gribble and Scott, 2002; Gribble *et al.*, 2003; Kelso, 1982; Zemlin, 1988). It also leads to inefficiency in model representations. For instance, to synthesize

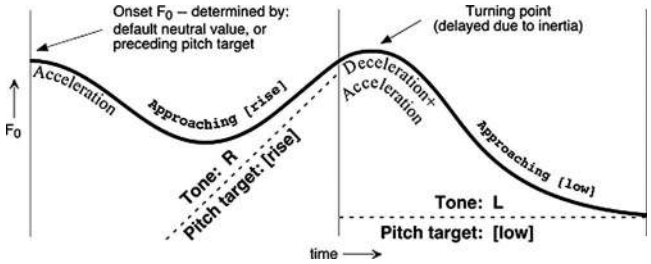

FIG. 1. Illustration of the theoretical TA model proposed in Xu and Wang (2001). In each syllable, demarcated by the vertical lines, surface F_0 (solid curve) asymptotically approaches the underlying pitch target (dashed line).

F_0 for each syllable, up to ten parameters are required. Although it is possible for the model to generate high quality resynthesis (Fujisaki *et al.*, 2005), implementing the communicative functions by summarizing these parameters together would be very complex.

The inadequacies of the existing models have motivated the proposal of the TA model (Xu and Wang, 2001). The model, as illustrated in Fig. 1, is based on the analysis of continuous acoustic data of Mandarin tone and intonation (Xu, 1997, 1998, 1999, 2001). The TA model assumes that observed F_0 contours are the outcome of implementing pitch targets which are linear functions that are either static (e.g., [low]) or dynamic (e.g., [rise]), as depicted by the dashed lines in Fig. 1. (Theoretically there may exist curvilinear targets in addition to linear ones, as explained in Xu (2005). But for the current implementation, there is no strong justification for including nonlinear targets.) The implementation of the pitch targets is synchronized with the syllable, i.e., starting at the onset of the syllable and ending at the offset of the syllable, based on evidence from acoustic data (Xu, 1998, 1999, 2001). In most cases, a tone is assumed to have only one pitch target (Xu and Wang, 2001). During each TA, the state of articulation depends not only on the discrepancy between the current state and the target but also on the final velocity and acceleration of the preceding syllable. In Fig. 1, for example, at the beginning of the second syllable, while the implementation of the [low] target has already started, F_0 is still rising due to the initial velocity and acceleration resulting from implementing the [rise] target in the first syllable. The influence of the preceding target, also known as carryover effect, would gradually decrease over time. Thus the state of articulation, as defined by F_0 height, velocity, and acceleration, is transferred from one syllable to the next at the syllable boundary. Such transfer of articulatory state is assumed to explain not only the well-known carryover assimilatory effects (Gandour *et al.*, 1994; Xu, 1997) but also the phenomenon of F_0 peak delay in both tone and nontone languages (Arvaniti and Ladd, 1995; Arvaniti *et al.*, 1998; de Jong, 1994; Ladd, 1983; Pierrehumbert and Steele, 1990; Prieto *et al.*, 1995; Silverman and Pierrehumbert, 1990; Xu, 2001).

However, the TA model alone can only describe the low-level articulatory process. A more complete framework is needed to link the articulatory mechanisms to the higher-level processes in speech. Xu (2005) proposed that TA is not limited to the realization of lexical tones but also serves as a

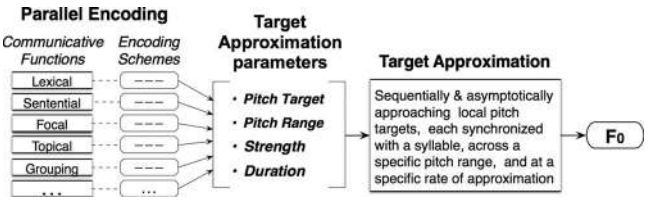

FIG. 2. A sketch of the PENTA model adapted from Xu (2005) with minor modifications. The stacked boxes on the far left represent individual communicative functions, which constitute the primary inputs to the model. They are parallel to each other with no hierarchical organizations, since the meanings they represent are independent of each other. The communicative functions are manifested through distinctive encoding schemes (second stack of boxes from left), which are either universal or language specific. The encoding schemes then specify the values of the TA parameters (middle block): pitch target, pitch range, strength, and duration, which are the control parameters of the TA model (Fig. 1) that simulates the articulatory process as syllable-synchronized sequential TA.

base mechanism for encoding other pitch related communicative meanings. That is, functions other than lexical tone also have to be encoded through the TA process, i.e., by manipulating the parameters of the process, namely, pitch range (which determines the pitch span of the targets), articulatory strength (which determines the speed of TA), and syllable duration (which assigns the amount of time for the approximation of each target). This is summarized as the parallel encoding and target approximation (PENTA) model, as illustrated in Fig. 2.

As can be seen in Fig. 2, the PENTA model assumes that speech prosody has to convey multiple communicative functions *in parallel* through individual encoding schemes. Though being abstract, these encoding schemes are always linked to specific functions. Thus it is through the TA process that a continuous link is maintained between multiple communicative functions and surface F_0 contours. Following this assumption, effective modeling of speech prosody can be achieved only if the encoding schemes of specific communicative functions are simulated. Although functional views of speech prosody are by no means new (e.g., Bailly and Holm, 2005; Bolinger, 1989; Hirschberg, 2002; Hirst, 2005; Kohler, 2005; Pierrehumbert and Hirschberg, 1990), PENTA, based on an articulatory-functional view, offers the implementational framework consisting of explicit mechanisms that directly link multiple communicative functions to the articulatory process of F_0 contour generation.

Both TA and PENTA, however, are conceptual models and thus need to be tested quantitatively. There have been previous attempts to quantify these models, including, in particular, Xu *et al.* (1999), but none of the earlier attempts has been fully satisfactory, as they are not able to simulate all the main mechanisms of target approximation revealed by previous acoustic analyses (Chen and Xu, 2006; Xu, 1997, 1998, 1999, 2001; Xu and Sun, 2002). In this paper we present the quantitative target approximation (qTA) model, which is the outcome of quantifying both TA and PENTA. We will first outline in Sec. II the basic assumptions about the articulatory mechanisms of F_0 production and how they are implemented in qTA. In Sec. III, we will explain how tone, lexical stress, and focus can be modeled based on our assumptions about the functional nature of pitch production in speech. Finally, we will report in Sec. IV the results of a series of experiments testing qTA through quantitative as well as perceptual evaluations.

II. MODELING BIOPHYSICAL MECHANISMS OF F_0 PRODUCTION

A. Background assumptions

In the following we present the background assumptions based on which the qTA model is developed. Although these assumptions are derived from recent research as just discussed, they are by no means treated as truth, but stated here so as to be explicit rather than hidden. Their validity will then be tested in the modeling experiments to be discussed subsequently. Some of the assumptions may seem to be too restrictive. But they help to reduce the degrees of freedom of the model, which is always desirable, other things being equal.

1. Vocal fold tension control as a third-order critically damped linear system

During phonation, the frequency of the vocal fold oscillation depends on the effective stiffness of the vocal folds which is directly proportional to vocal fold tension (Fujisaki, 2003; Titze, 1989) and also, though somewhat less directly, to subglottal pressure (Monsen *et al.*, 1978; Ohala, 1978; Titze, 1989). As suggested by the body-cover concept (Hirano, 1974), effective stiffness is related to the activation of the CT and the thyroarytenoid muscles, which are antagonistic to each other. The differential muscular control of the vocal folds generates two major muscle actions: increasing or decreasing their surface tension, which in turn raises or lowers F_0. Thus F_0 raising is done not only by the contraction of the CT muscle but also by the simultaneous antagonistic contraction of the thyroarytenoid muscle. Likewise, F_0 lowering is done not only by the reduction in CT contraction but also by a simultaneous thyroarytenoid contraction, which shortens the vocal folds. It is also known that the production of very low F_0 involves the extrinsic laryngeal muscles such as the sternohyoids, sternothyroids, and omohyoids (Erickson, 1976; Erickson *et al.*, 1995; Hallé, 1994). Thus the F_0 raising and lowering actions would be further aided by the contraction of these extrinsic laryngeal muscles in the lower pitch range of a speaker. A biomechanical system of the tension force controlled by antagonistic muscles that transfer energy back and forth within the system can be represented by an Nth-order linear system (Palm, 1999), where N is the number of the energy-storage elements. The qTA model is thus configured as an Nth-order linear system. Physiologically, as discussed earlier, there are at least two major antagonistic muscle forces controlling the vocal fold tension and various minor influences from the extrinsic laryngeal muscles and the subglottal pressure. Thus the model should be at least second order. This aspect of qTA is, to a large extent, similar to the CR model for F_0 control (Fujisaki, 1983; Fujisaki *et al.*, 2005) and the task dynamic (TD) model for the control of the segmental aspect of speech (Saltzman and Kelso, 1987; Saltzman and Munhall, 1989). Neverthe-

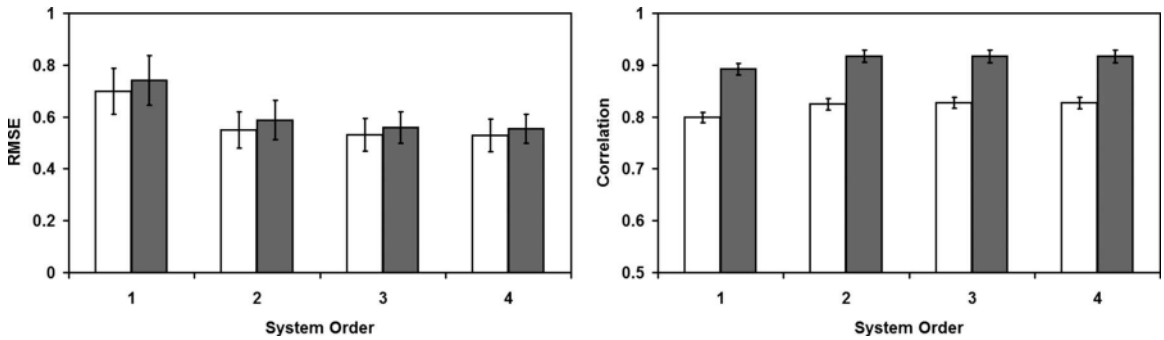

FIG. 3. Average RMSE (left) and correlation (right) of resynthesis results comparing between damping conditions and model order. White bars indicate results from the overdamped system and dark bars indicate results from the critically damped system. Vertical lines show standard errors of the mean.

less, qTA is different from CR and TD in a number of nontrivial ways, which will be highlighted as our description of qTA proceeds.

*N*th-order linear systems can have many possible damping behaviors. For instance, a second-order linear system can respond to externally imposed forces with three possible damping behaviors: underdamped, overdamped, and critically damped. Of these, critical damping can be viewed as the borderline condition between underdamping and overdamping. The responses of an underdamped second-order system have a shorter rise time compared to those of an overdamped system, but they also manifest oscillation. In contrast, the responses of an overdamped system have a longer rise time with no oscillation. In speech, F_0 movements seem to be goal oriented once the underlying functional components (e.g., tone, stress, or focus) are taken into consideration (Chen and Xu, 2006; Gandour *et al.*, 1994; Wong, 2006; Xu, 1997, 1999). In Mandarin, for example, an F_0 drop after the high tone is not for the sake of returning to a baseline or a neutral pitch register but to approach the lower pitch onset of another tone (e.g., low, rising, or the neutral tone). Similar evidence has also been found in English (Xu and Xu, 2005). This suggests that the control mechanism of F_0 regulation for speech is nonoscillatory, which means that the system is more likely to be overdamped than underdamped. An overdamped system has previously been suggested for supralaryngeal articulation (Fujisaki, 1974; Lindblom, 1983). Thus the candidate damping behaviors of the F_0 production model can be either overdamped or critically damped, although the natural process is more likely to be overdamped.

Although an overdamped linear system with an order equaling to the number of factors that antagonistically control the vocal fold tension would be a desirable choice, it is not mathematically efficient because of its complexity. Rather, a simpler model that can still generate F_0 contours with sufficient accuracy would be preferred. To determine the suitable order and damping behavior, a pilot test was conducted by resynthesizing the F_0 contours in a Mandarin corpus (Xu, 1999) and measuring the root mean square error (RMSE) and Pearson's correlation coefficient (henceforth, correlation) between the synthesized and original F_0. The detail of the corpus will be later discussed in detail in Sec. IV A while the parameter extraction method will be explained in Sec. II C. For each syllable, the model parameters, including pitch target and approximation rate, were estimated and then used to synthesize the F_0 contour. Afterward, the RMSE and correlation were calculated. The mean and confidence interval of RMSE and correlation were then derived from the mean of RMSE and correlation of each speaker. As shown in Fig. 3, although RMSE of the critically damped system is slightly lower than that of the overdamped system, their differences are not significant [$F(1,31)=0.35$, $p=0.556$, with damping condition and system order as independent factors] while the order significantly affects the error [$F(3,31)=2.80$, $p=0.047$]. In contrast, for the correlation between the synthesized and original F_0, the critically damped system is significantly better than the overdamped system [$F(1,31)=132.92$, $p<0.001$] while the effect of system order is still significant [$F(3,31)=2.90$, $p=0.042$]. Moreover, for a general *N*th-order linear system, the critically damped system has only one time-constant parameter while an overdamped system has *N* time-constant parameters (see Table I which will be explained next). Mathematically, a critically damped system is therefore simpler and possibly more accu-

TABLE I. Average rates of TA (λ, as will be explained in Sec. II B 2) and their 95% confidence intervals as functions of damping condition and model order. The number of λ values depends on the model order in an overdamped system while there is only one λ value in a critically damped system regardless of order.

Order	Overdamped (s^{-1})					Critically damped (s^{-1})
1	32.7 ± 4.3					20.5 ± 2.6
2	33.6 ± 1.3	52.3 ± 1.2				30.0 ± 2.5
3	37.9 ± 1.0	55.5 ± 1.3	68.3 ± 1.8			44.6 ± 2.9
4	35.4 ± 1.1	54.9 ± 0.9	69.1 ± 1.6	71.3 ± 1.5		49.4 ± 1.9
5	36.1 ± 1.5	59.2 ± 1.2	66.4 ± 1.5	68.6 ± 1.9	75.2 ± 1.2	50.9 ± 3.2

rate than an overdamped system. Also, as seen in Fig. 3, while the system performance is improved from second to third order in terms of both RMSE and correlation, there is little improvement from third to fourth order. Using an exponential regression to determine the lowest order at which RMSE is reduced to be within 5% from the steady state ($\lceil 3 \times \text{time constant} \rceil = \lceil 3 \times 0.8 \rceil = 3$, where $\lceil x \rceil$ is a ceiling function), we found that it is sufficient to simulate F_0 production with a third-order critically damped linear system. Table I shows the average rates of TA of both overdamped and critically damped systems with different system orders. The values in the table indicate how fast F_0 approaches the desired pitch target in terms of both slope and height when the accuracy of estimation was optimal for each system order. Last but not the least, an added advantage of a third-order system over a second-order one is that it guarantees smoothness across syllable boundaries in terms of not only F_0 level and velocity but also acceleration, which better simulates the cross-boundary state transfer assumed in the TA model.

2. Sequentiality and syllable synchronization

In the TA model, it is assumed that the most local components of tone and intonation are strictly sequential in articulation and are fully synchronized with the syllable (Xu and Wang, 2001). This assumption is based on evidence from empirical research that the F_0 movement toward an underlying target starts from syllable onset rather than from voice onset even in syllables with a voiceless initial consonant (Xu and Xu, 2003 for Mandarin; Xu and Wallace, 2004 for English; Wong and Xu, 2007 for Cantonese) and that the target approaching movement ends at syllable offset rather than at vowel offset when the syllable has a coda consonant (Xu, 1998, 2001 for Mandarin; Wong and Xu, 2007 for Cantonese). There has also been evidence that syllable-based F_0 modeling is not only feasible (Black and Hunt, 1996; Ross and Ostendorf, 1999) but also superior to accent-based modeling that ignore syllable boundaries (Sun, 2002). Furthermore, although most existing F_0 models do not hold this assumption (e.g., Fujisaki *et al.*, 2005; Kochanski and Shih, 2003; Taylor, 2002; van Santen and Möbius, 2000; but see Fujimura, 2000 for a syllable-based gestural-organization model), at least two modeling efforts have generated evidence for synchronization of the syllable with tonal units (Kochanski *et al.*, 2003) or tonal commands (Gu *et al.*, 2007). It is worth reiterating here that what is sequential and syllable synchronized is the underlying target, namely, the equivalent of the dashed lines in Fig. 1, rather than any surface F_0 event such as turning point, which is apparently delayed beyond syllable 1 in Fig. 1. This is a critical point where our assumption differs from the conclusions of many other studies that suggest variable alignment based on surface F_0 events such as peaks, valleys, and elbows (e.g., Arvaniti *et al.*, 1998; Atterer and Ladd, 2004; Chen *et al.*, 2004; Kohler, 2005).

From the perspective of modeling, sequentiality and syllable synchronization robustly reduce the degrees of freedom for the control of the TA process. That is, the implementation of a tone always starts from the onset of the syllable and ends at the offset of the syllable, as shown in Fig. 1. Nevertheless, the state of articulation, as specified by pitch level, velocity, and acceleration, is transferred from one syllable to the next at the syllable boundary. This differs from the CR model which assumes only the transfer of displacement across the executions of adjacent commands. There is no transfer of velocity or acceleration as far as we can see from the published descriptions of the model (Fujisaki, 1983, 2003; Fujisaki *et al.*, 2005).

Another important aspect of sequentiality is the assumption that all movements unidirectionally approach one target or another, with no return phases to a baseline or a neutral position. Such return phases are either obligatory or optional in other models based on a damped linear system (e.g., Fujisaki *et al.*, 2005; Saltzman and Munhall, 1989).

B. Modeling

1. Pitch target

In qTA, a pitch target is defined as the underlying goal of the local tonal or intonational component. It is a forcing function, representing the joint force of the laryngeal muscles that controls vocal fold tension. Based on the observation of the surface F_0 contours in continuous speech (Xu, 1997, 1999) and the theoretical conceptualization of the TA model (Xu and Wang, 2001), a pitch target can be represented by a simple linear equation

$$x(t) = mt + b, \tag{1}$$

where m and b denote the slope and height of the pitch target, respectively. Since the implementation of the pitch target is local to the host syllable, the time, t, is relative to the onset of the syllable.

There are two types of targets: static, e.g., [high], and dynamic, e.g., [rise]. In a static target, the slope m equals zero, while in a dynamic target m is either positive or negative. The empirical studies of Mandarin tones (Xu, 1997, 1999, 2001; Xu and Wang, 2001; Chen and Xu, 2006) suggest that high (H), low (L), and neutral (N) tones can be represented by static targets: [high], [low], and [mid], while rising (R) and falling (F) tones can be represented by dynamic targets: [rise] and [fall]. As demonstrated in Xu and Wang (2001), the slopes of the dynamic targets are essential to the dynamic tones like R and F because their F_0 variability at different speech rates cannot be adequately simulated by sequences of static targets such as [low+high] for [rise] or [high+low] for [fall]. Moreover, recent studies of English intonation suggest that an unstressed syllable may be represented by a static target [mid], while a stressed syllable may have either a static or dynamic target depending on a number of lexical and postlexical factors (Xu and Xu, 2005).

2. F_0 realization

The control of vocal fold tension in qTA is implemented through a third-order critically damped linear system. Generally, the response of the linear system consists of two parts: forced response and natural response. The forced response is the output of the system when it reaches the desired state, as is assumed in the TA model. The natural response is the

transient in the transition from the current articulatory state to the desired state represented by the pitch target. This transient effect diminishes over time. For a third-order critically damped system, the total response is in the form of

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

where the first term, $x(t)$, is the forced response which is the pitch target itself and the second term, the polynomial and the exponential, is the natural response. $f_0(t)$ is the complete form of the fundamental frequency in hertz. The model thus has three parameters, m and b which specify the pitch target and λ which represents the rate of TA. The transient coefficients c_1, c_2, and c_3 are determined jointly by the initial conditions and the target of the articulatory process. The initial conditions are the initial state of the dynamic F_0 movement, consisting of initial F_0 level, $f_0(0)$, initial velocity, $f_0'(0)$, and initial acceleration, $f_0''(0)$. Solving the systems of linear equations determined from the initial conditions, the transient coefficients can be expressed in the following formulas:

$$c_1 = f_0(0) - b, \quad (3)$$

$$c_2 = f_0'(0) + c_1\lambda - m, \quad (4)$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2. \quad (5)$$

Figure 4 shows the F_0 responses as each model parameter varies. Figures 4(a) and 4(b) demonstrate the F_0 contours when the first and second targets vary. Figures 4(c) and 4(d) illustrate the F_0 contours when λ varies for different combinations of dynamic and static targets. Other things being equal, F_0 approaches the target more rapidly as λ increases. As a result, the shape of the F_0 contour varies greatly with the change in the λ value. Peak delays can be clearly observed in Fig. 4(d) as λ of the second syllable varies. The approximation rate represented by λ is therefore an important parameter of the TA process, as has been suggested in previous empirical research (Chen and Xu, 2006; Xu and Xu, 2005).

C. Automatic parameter extraction

Parameter extraction was done with an automatic analysis-by-synthesis optimization algorithm, as illustrated in Fig. 5. The purpose of the automatic parameter extraction is to train the model based on available data so that we can both resynthesize the F_0 contours in the training data and synthesize novel F_0 contours. The algorithm reads the training data and parameter constraints, then automatically varies the parameter values in the specified search space, and finally adopts the parameter set with the lowest sum square error between the synthesized and original F_0 contours. The optimization is conducted one syllable at a time according to the sequentiality and syllable synchronization assumptions.

In traditional acoustic synthesis by rule, the parameters used are based on human understanding of the speech production process. These parameters can be obtained either from measurements of acoustic or articulatory data or from analysis by synthesis based on the proposed model (Klatt, 1987). For models that represent F_0 contours independent of physical mechanisms, the parameters are usually data driven based on real speech corpora. However, for models that simulate underlying mechanisms of certain processes, the analysis-by-synthesis method is typically used to estimate the model parameters (Fujisaki *et al.*, 2005; Kochanski and Shih, 2003; Mixdorff, 2000). This is because it is difficult, and sometimes impossible, to construct inverse algorithms for the production models. With an analysis-by-synthesis framework, parameter estimation can be done for any system regardless of its reversibility.

In qTA, for each syllable, the parameter λ represents its assigned level of articulatory effort. It corresponds to the rate of TA and is inversely proportional to the time constant of the approximation process. The time constant is the time relative to the onset of the current TA. Figure 6 shows computed percentages of F_0 approximation to a static target with different time constant. They are calculated by substituting t in Eq. (2) with the number that is a multiple of the time constant. A higher λ indicates that F_0 will reach the target faster. It requires more than five time constants to achieve 90% of the target. It should be noted that λ is not directly equal to the speed of F_0 movement because the same λ may result in different F_0 speed depending on b and m of the underling pitch target. In the present study, λ is allowed to vary from 0 to 120 s^{-1}, which correspond to the time to fully reach a pitch target in a very short syllable (≈70 ms). This search range is also consistent with the computed percentage of TA shown in Fig. 6.

For the other two variable parameters, namely, m and b, both specifying the pitch target, we also imposed restrictions based on our understanding of the nature of the targets as discussed earlier. For each target we specified a search space for m: zero for [high], [low], and [mid], positive for [rise], and negative for [fall]. Also for each target, we restricted the search space of b of each syllable to be within a small range (±20 Hz). The center point of this search space is around the predicted y-intercept value which is the difference between the final F_0 value and the expected excursion, i.e., target slope × syllable duration. This is because, based on previous findings, syllable offset is where surface F_0 becomes closest to the target (Xu, 1997, 1999; Xu and Wang, 2001). In cases voicing stops before the end of the syllable, the last available F_0 value and its corresponding time were used instead.

Note that these restrictions are based on previous empirical findings about tone production rather than being arbitrary, as discussed earlier in this section. Once imposed, they help to significantly reduce the possibility of the searches being stuck at local minima, thus allowing the training to be fully automatic. Table II shows the constraint violation rates and relative changes in standard deviation of estimated model parameters. The constraint violation rate is the percentage of the number of parameter estimations that fall outside the ranges of parameter constraints. The relative change in standard deviation is calculated as the relative difference between the standard deviation obtained with and without parameter constraints. The results in Table II were derived from the Mandarin dataset which will be discussed later in Sec. IV A. As shown in Table II, the most influential con-

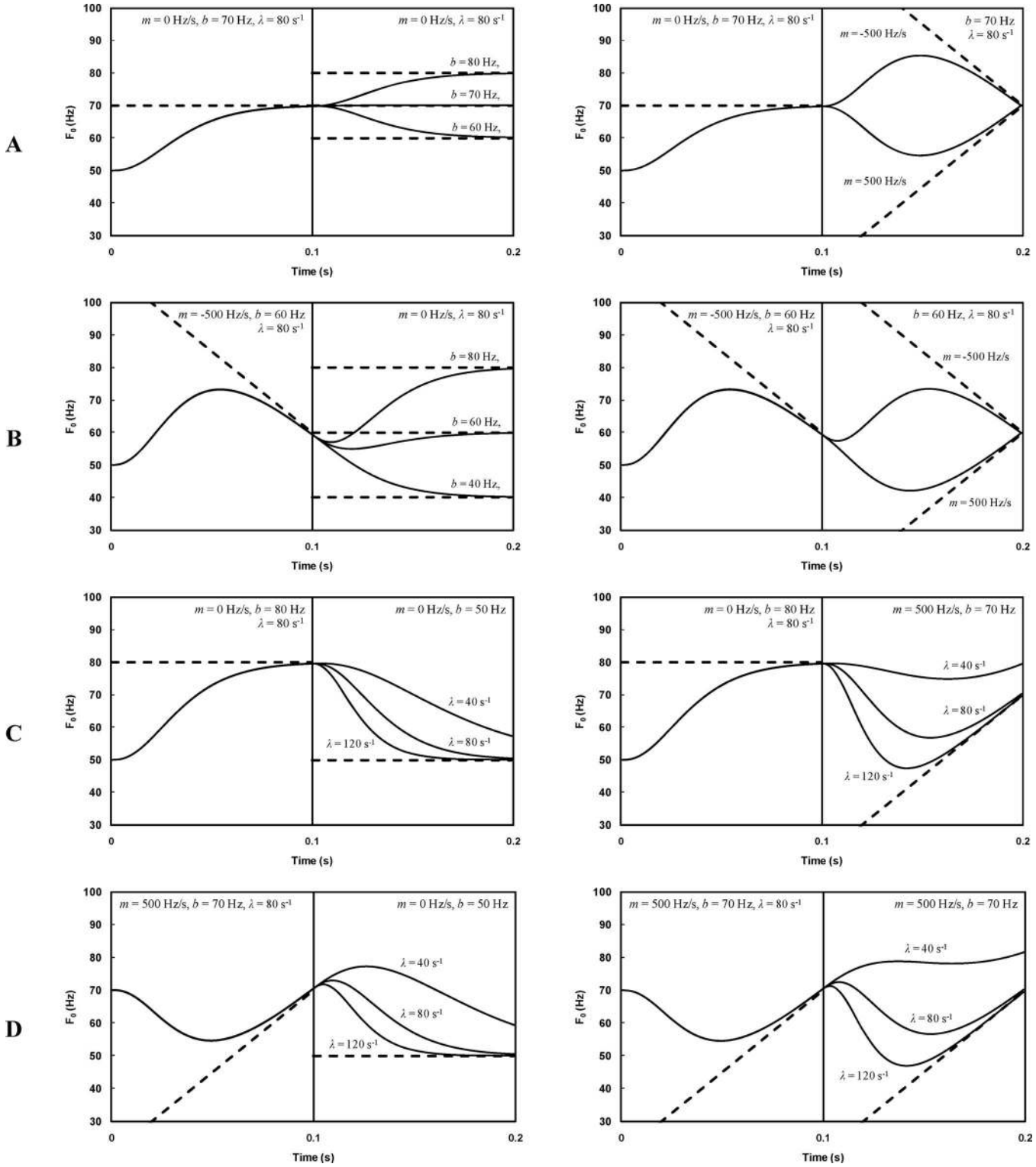

FIG. 4. Examples of F_0 contours generated by the qTA model with varying values of m, b, and λ. The dashed lines indicate the underlying pitch targets which are linear functions of m and b. The vertical lines show the syllable boundaries through which the articulatory state propagates.

straint is the m-constraint because discarding it results in the highest constraint violation rate (52.35%). Most of the m violations occurred in the H and L tones, for which the constraint $m=0$ is extremely easy to violate, as hardly any surface F_0 contour is fully flat by the end of the syllable with those tones. The right half of Table II shows that removing the constraint on m also affects the variation in λ (81.22% greater standard deviation), while the b-constraint moderately affects variations in both b and λ. The least influential is the λ-constraint which mainly affects the variation in λ itself and affects those of other parameters only slightly. Interestingly, the only parameter that has been heavily affected by constraints other than its own is λ. This indicates that the estimation of λ is quite dependent on the estimation of the pitch target parameters.

An example of the F_0 contour synthesized with param-

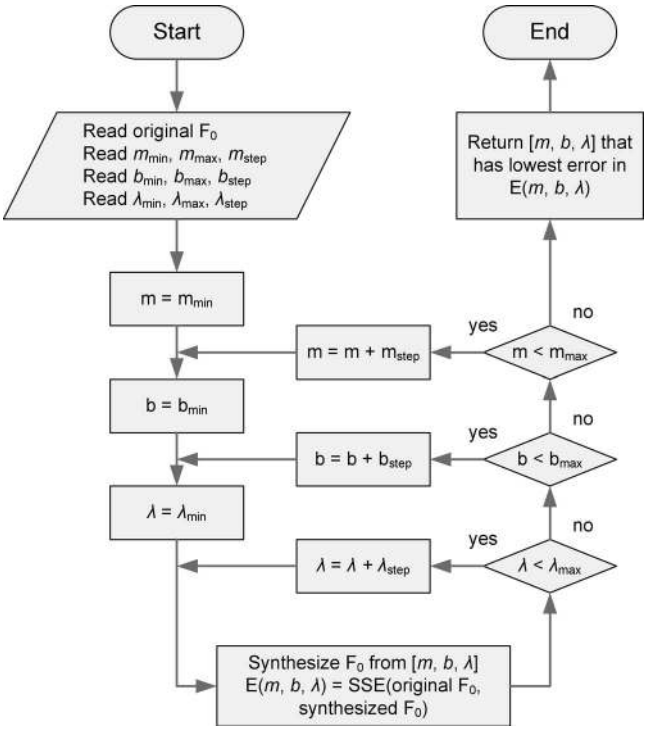

FIG. 5. A flowchart of automatic parameter extraction of qTA model. The algorithm optimizes for the suitable model parameters that, when implemented by the qTA model, generate F_0 contours that closely approximate the original ones.

eters resulting from the automatic extraction is shown in Fig. 7. The synthesized F_0 closely approximates the original one, as will be later shown in the low average RMSE in the testing on all speakers.

III. MODELING TONE AND FOCUS AS COMMUNICATIVE FUNCTIONS

As suggested by the PENTA model discussed in Sec. I, speech prosody conveys multiple communicative functions in parallel. Each function is represented by a unique encoding scheme, specified in terms of one or more of the TA parameters: pitch target, pitch range, strength, and duration, as illustrated in Fig. 2. Effective quantitative modeling of

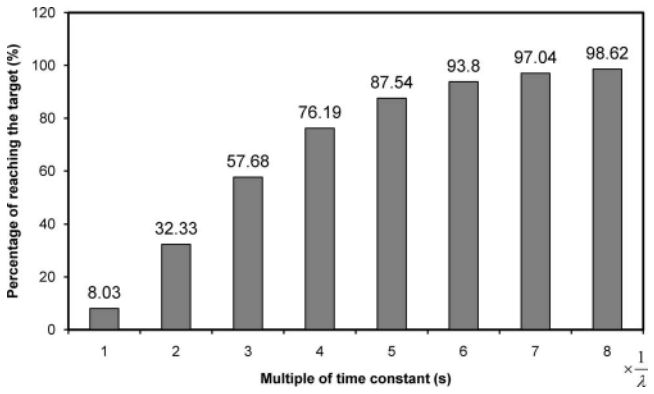

FIG. 6. Percentage of F_0 approximation to a static target as a function of time constant. These percentages were calculated by substituting t in Eq. (2) with different multiples of the time constant.

TABLE II. Constraint violation rates and relative changes in standard deviation when removing each parameter constraint in the automatic parameter extraction.

Constraint on	Constraint violation rate (%)	Relative change in standard deviation (%)		
		$\Delta\sigma_m$	$\Delta\sigma_b$	$\Delta\sigma_\lambda$
m	52.35	200.82	−0.67	81.22
b	17.15	−4.43	13.75	10.84
λ	7.21	1.74	−0.71	20.22

speech prosody, based on this view, can be achieved only if individual communicative functions are simulated.

To formalize encoding schemes of the communicative functions, we define the *parametric vector* of the jth syllable as a set of qTA model parameters, $\mathbf{p}_j=[m_j, b_j, \lambda_j]^T$. The F_0 of the sentence can be formed by consecutively executing a series of parametric vectors. A *prosodic vector* of length N represents the prosody of the sentence with N syllables. It can be expressed in the form of a series of N parametric vectors, $\mathbf{s}=\{\mathbf{p}_1, \mathbf{p}_2 \cdots \mathbf{p}_N\}$. This vector is the input to the TA process. It is possible for communicative functions to be represented by either parametric vectors or prosodic vectors, depending on the temporal domain of the function.

Three communicative functions are directly modeled in the present study, lexical tone in Mandarin, lexical stress in English, and focus in both languages, as explained next. Two other communicative functions, namely, new topic and sentence modality (Swerts, 1997; Lehiste, 1975; Liu and Xu, 2005; Wang and Xu, 2006), are only indirectly modeled as positional dependencies of the functions. Figure 8 is a block diagram of qTA and the communicative functions implemented in the present study. The input communicative functions are functional specifications which can be tone/stress symbols for the tone/stress function or focus position for the focus function. The generated parametric vector from the tone/stress function is additively combined with the focus adjustment vector from the focus function. This combination results in the focus-adjusted parametric vector which is the output of the parallel encoding process and also the input to the TA process.

A. Lexical tone and lexical stress

For Mandarin, the most local functional components are the lexical tones. Given a symbolic input tone x, the tone function generates the parametric vector $\mathbf{p}$:

$$\mathbf{p} = \text{tone}(x), \tag{6}$$

where the tone category x is N, H, R, L, or F. The tone function returns a parametric vector for tone x. Thus each tone requires one parametric vector to represent it. During training, the parameter values of a tone are derived by averaging the parametric vectors extracted from all individual occurrences of that tone.

For English, the most local pitch component is lexical stress. Unlike in Mandarin, however, pitch values related to lexical stress in English are assigned postlexically (Ladd, 1996; Liu and Xu, 2005; Pierrehumbert, 1980; Xu and Xu, 2005). Xu and Xu (2005) showed that unstressed syllables in

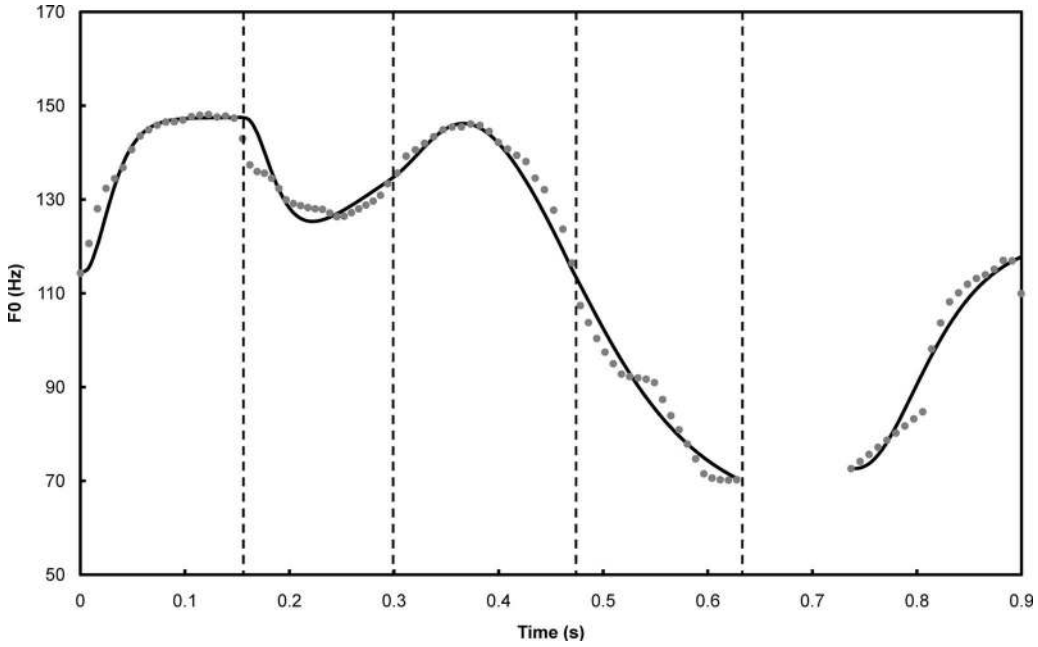

FIG. 7. An example of resynthesized F_0 for the tone sequence HRFLH. The solid line represents the synthesized F_0 contour while the gray dotted line indicates the original F_0 contour. The dashed vertical lines show the syllable boundaries. The discontinuity of F_0 at the beginning of the last syllable is due to the voiceless initial stop consonant [t] in the last syllable.

English, just like the neutral tone in Mandarin (Chen and Xu, 2006), are assigned specific pitch targets rather than being targetless. Thus for English, every syllable is assigned a pitch target regardless of its stress status. The functional representation of lexical stress is

$$\mathbf{p} = \text{stress}(x), \tag{7}$$

where x represents the stress category which is either *stressed* or *unstressed*.

B. Focus

Focus is a discourse function to highlight a particular piece of information against the background information (Bolinger, 1986; Gussenhoven, 2007; van Heuven, 1994; Ladd, 1996; Xu and Xu, 2005). In many languages, including English and Mandarin, focus expands the pitch range of the focused syllable(s) and compresses the pitch range of the postfocus syllable(s) (Botinis *et al.*, 2000; Hasegawa and Hata, 1992; Krahmer and Swerts, 2001; Mixdorff, 2004; Rump and Collier, 1996; Xu, 1999; Xu and Xu, 2005). Based on these findings, syllables in each sentence can be grouped into maximally four regions: prefocus, on focus, postfocus, and final focus, whichever is applicable. Little pitch range adjustment occurs in the prefocus region. Pitch range of the on-focus region is expanded, while that of the postfocus region, which includes all syllables up to the end of the sentence, is compressed. Final focus is treated separately because it has been found to have a pitch range larger than that of neutral focus but smaller than that of nonfinal focus in both Mandarin and English (Xu, 1999; Xu and Xu, 2005). For a sentence with no narrow focus, its entirety is treated as prefocus.

Computationally, focus is treated as an adjustment func-

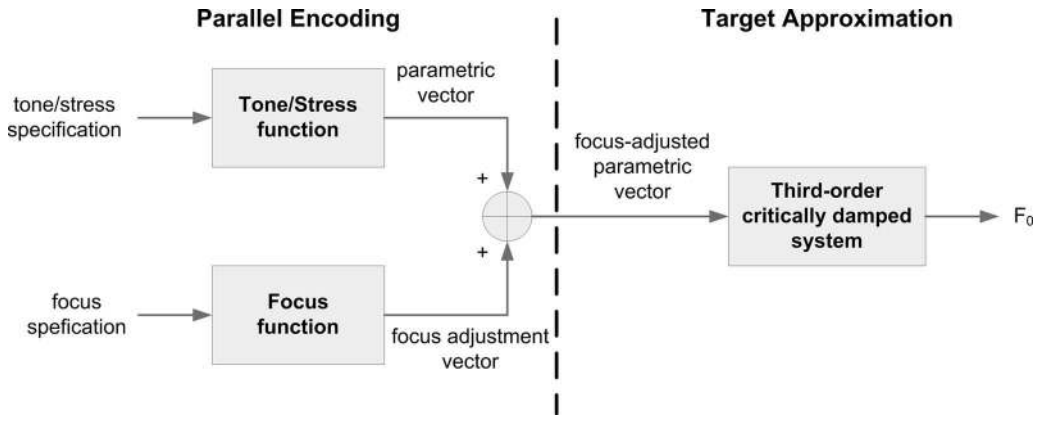

FIG. 8. A block diagram of the qTA model as the quantitative counterpart of the PENTA model. The dashed vertical line separates the parallel encoding, which mainly generates and manipulates the parametric vectors, from the TA, which implements the parametric vectors and synthesizes F_0.

tion. It maps the given prosodic vector **s** with length N and focus position **K** to the output prosodic vector $\hat{\mathbf{s}}$.

$$\hat{\mathbf{s}} = \text{focus}(\mathbf{s}, \mathbf{K}), \tag{8}$$

where **K** is a set of positive integers including one but less than or equal to N. Each element in the set **K** indicates the position of the syllables of the word under focus. If the focused word is monosyllabic, **K** has only one element. Focus encoding is done by adjusting the input prosodic vector according to the trained focus adjustment parameters. For instance, the implementation of medial focus at the Mth syllable ($\mathbf{K}=\{M\}$) would be

$$\hat{\mathbf{s}} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M + \Delta\mathbf{p}_{\text{on}}, \mathbf{p}_{M+1} + \Delta\mathbf{p}_{\text{post}}, \ldots, \mathbf{p}_N + \Delta\mathbf{p}_{\text{post}}\}, \tag{9}$$

where $\Delta\mathbf{p}_{\text{on}}$ and $\Delta\mathbf{p}_{\text{post}}$ denote the focus adjustment vectors of on- and postfocus regions, respectively. Here, no adjustment is made before focus. The on-focus parametric vector is changed by the on-focus adjustment which expands the pitch range of the pitch target. The remaining parametric vectors are modified by the postfocus adjustment which compresses and lowers the pitch range of the pitch targets. Similarly, the implementation of final focus at the last syllable ($\mathbf{K}=\{N\}$) would be

$$\hat{\mathbf{s}} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N + \Delta\mathbf{p}_{\text{final}}\}, \tag{10}$$

where $\Delta\mathbf{p}_{\text{final}}$ denotes the focus adjustment vectors of final-focus regions. Only the parametric vector of the last syllable is adjusted in this case. If the focused word consists of two syllables, the implementation of the initial focus of the first two syllables ($\mathbf{K}=\{1,2\}$) would be

$$\hat{\mathbf{s}} = \{\mathbf{p}_1 + \Delta\mathbf{p}_{\text{on}}, \mathbf{p}_2 + \Delta\mathbf{p}_{\text{on}}, \mathbf{p}_3 + \Delta\mathbf{p}_{\text{post}}, \ldots, \mathbf{p}_N + \Delta\mathbf{p}_{\text{post}}\}. \tag{11}$$

In training the focus function, the extracted parametric vectors from the training corpus are divided into four focus regions, pre-, on-, post-, and final focus. The parametric vectors in the prefocus regions are averaged together according to tone or lexical stress and treated as the reference parametric vector for further calculation in other focus regions. For other focus regions, the focus adjustment vectors are calculated by averaging the differences between the parametric vectors in on-, post-, or final-focus regions and the reference parametric vectors, respectively.

$$\bar{\mathbf{p}} = \frac{1}{N}\sum_i \mathbf{p}_i, \quad \mathbf{p}_i \in \text{prefocus set}, \quad N = \text{total number of } \mathbf{p}_i \text{ in prefocus set}, \tag{12}$$

$$\Delta\mathbf{p}_{\text{floc}} = \frac{1}{M}\sum_j (\mathbf{p}_j - \bar{\mathbf{p}}), \quad \mathbf{p}_j \in \text{floc focus set}, \quad M = \text{total number of } \mathbf{p}_j \text{ in floc focus set}, \tag{13}$$

where floc denotes three possible focus locations including on, post, and final. Thus, for example, the total number of parametric vectors required for tone and focus representation in Mandarin is 16:4 prefocus parametric vectors and 12 focus adjustment vectors. For English, the total number of parametric vectors is 8:2 prefocus parametric vectors and 6 focus adjustment vectors.

IV. EXPERIMENTAL EVALUATION

A. Corpora

Two datasets were used in the experiments for testing the ability of the qTA model to simulate tone and intonation in Mandarin and English. The Mandarin dataset was originally collected for a study of tone and focus (Xu, 1999). It consists of 3840 five-syllable utterances recorded by four male and four female Mandarin speakers. Figure 9 illustrates the cross-speaker and cross-gender variations in F_0 contours in the tone sequence HRFHH produced by eight speakers in different focus conditions. The difference between genders can be easily observed in the two nonoverlapping clusters in each panel. Within-gender variability can also be clearly seen in the spread of the contours within each cluster.

The words in the sentences and the corresponding tones in this dataset are shown in Table III. In each utterance, the first two and last two syllables are disyllabic words while the third syllable is a monosyllabic word. The first and last syllables in each sentence always have the H tone while the tones of the other syllables vary depending on the position: H, R, L, or F in the second syllable, H, R, or F in the third syllable, and H or L in the fourth syllable. Since the dataset was originally designed for studying tone and focus, each sentence has four focus conditions: no focus, initial focus, medial focus, and final focus. Thus, there are totally 96 variations in tone and focus.

The English dataset was originally collected for a study of the effect of focus on English intonation (Xu and Xu, 2005). It consists of 1176 short declarative utterances recorded by four male and four female American English speakers. A list of the sentences in this dataset is shown in Table IV. Each sentence is said in one of four focus conditions: no focus, sentence-initial focus, sentence-medial focus, and sentence-final focus, and also in one of two speaking rates: normal and fast. For each focus condition and speaking rate, the sentences were repeated seven times. The dataset is further divided into three sentence groups, which differ in the position of the target word. In total, there are 42 combinations of word, focus condition, and speaking rate.

Figure 10 shows examples of F_0 extracted from both datasets and then averaged across speakers. The top left panel shows examples of Mandarin with tonal variations at the second syllable. The bottom left panel shows the effects of focus variation on F_0 when the tones are fixed. The top right and bottom right panels show examples from two different sentences in English with different types of focus.

B. Parameter analysis

Using the automatic analysis-by-synthesis optimization algorithm described in Sec. II C, we extracted the prosodic vectors from each sentence in the two datasets. After that, we derived function-specific parametric vectors, consisting of the model parameters m, b, and λ, for different communicative functions. For Mandarin, parameters for the tone and

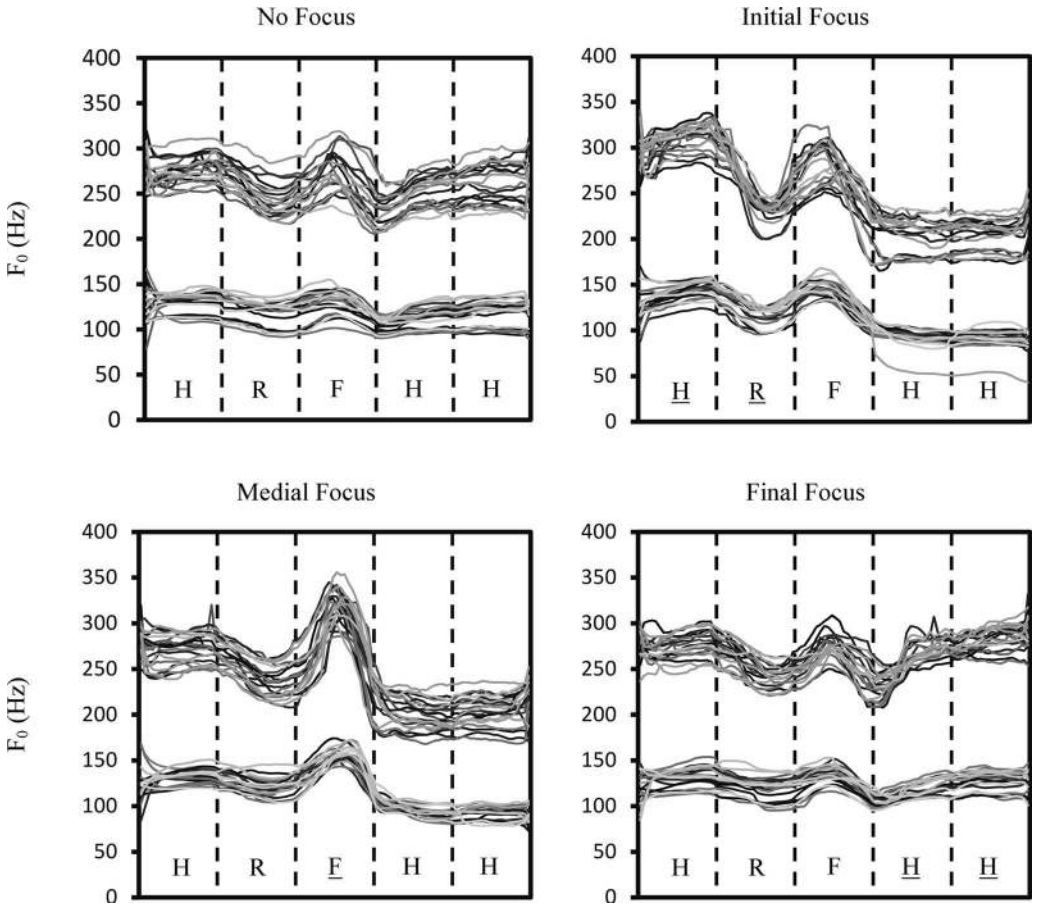

FIG. 9. F_0 contours of the HRFHH tone sequence spoken by eight speakers in different focus conditions in the Mandarin corpus. The focused syllable is denoted as the underlined letter. Each panel contains 40 F_0 contours (five repetitions by each speaker). The vertical dashed lines mark the syllable boundaries. In each panel the two clusters are from the female and male speakers, respectively.

focus functions were extracted. For English, parameters for the stress and focus functions were extracted.

In intonation modeling, F_0 estimation algorithms typically represent F_0 in hertz scale, especially those that model surface F_0 directly (Hirst and Espesser, 1993; Pierrehumbert, 1981; Taylor, 2000; van Santen and Möbius, 2000). The hertz scale is also commonly used for F_0 display in most speech analysis applications. However, there is evidence that pitch operates in speech on a logarithmic scale in both perception (Nolan, 2003) and production (Fujisaki, 2003). Thus it is inappropriate to pool data across different speakers in hertz during modeling as they may have very different pitch spans depending on their pitch ranges. The current version of qTA model thus uses the semitone (st) scale for parameter representation while the hertz scale is used only for displaying F_0 so that the surface contours more directly reflect what is commonly seen. For speaker normalization purposes, the semitone scale is used for normalizing pitch ranges across speakers, and utterance-onset F_0 is used as a baseline to normalize speaker differences in terms of average F_0 level. For cross-speaker normalization, the extracted pitch target parameters are converted from hertz scale to semitone scale before averaging within each designated communicative function.

1. Mandarin

Table V shows the means and confidence intervals of the extracted parametric vectors of the tone function. They are extracted from data in prefocus regions, which also include all the words in sentences with no narrow focus, as explained earlier. The b values are measured relative to the initial F_0 of

TABLE III. Words and corresponding tone patterns of the Mandarin dataset (Xu, 1999). H, R, L, and F represent high, rising, low, and falling tones, respectively. The numerals at the end of each syllable also represents the tones: 1, 2, 3, and 4 for H, R, L, and F, respectively.

Word 1			Word 2			Word 3		
HH	mao1 mi1	"kitty"	H	mo1	"touches"	HH	mao1 mi1	"kitty"
HR	mao1 mi2	"cat-fan"	R	na2	"takes"	LH	ma3 dao1	"sabre"
HL	mao1 mi3	"cat-rice"	F	mai4	"sells"			
HF	mao1 mi4	"cat-honey"						

TABLE IV. A list of sentences in the English dataset (Xu and Xu, 2005).

Word 1	Word 2	Word 3	Word 4	Word 5
Lee/Nina/Lamar/ Emily/Ramona	May	Know	My	Niece
Lee		Lure/mimic/ minimize		Niece
Lee		Know		Niece/nanny/ mummy

TABLE V. Means and confidence intervals of extracted parametric vectors of the tone function obtained from the Mandarin dataset. Because different speakers have different average F_0, the utterance-onset F_0 is subtracted from b, the pitch target height, so that the values of b in the table are relative to the utterance-onset F_0.

Tone	m (st/s)	b (st)	λ (s^{-1})
H	0	0.0 ± 1.0	54.5 ± 5.4
R	93.4 ± 3.4	-2.2 ± 1.1	40.7 ± 3.8
L	0	-8.9 ± 0.6	34.1 ± 5.0
F	-106.4 ± 3.0	-2.5 ± 1.3	39.3 ± 3.2

each sentence. The model parameters in Table V show clearly distinct target values for the tones for both male and female speakers. Note that the positive and negative m for the R and F tones would have to be each represented by a combination of both positive and negative commands in the CR model (Fujisaki *et al.*, 2005), which demonstrates the greater simplicity of the qTA than the CR model in representing the tone function. The smaller absolute value of m in the R tone than in the F tone is consistent with the empirical finding that the maximum speed of pitch rises is lower than that of pitch drops (Xu and Sun, 2002). It should be noted that, for Tables V–IX, the confidence intervals of m and Δm for both H and L tones equal zero because those tones are assumed to have static targets. Note also that the small confidence intervals here and in Tables VI–X are partially attributable to the restrictions placed on the search space. Some of the confidence intervals may become larger once the restrictions are removed, as shown in Table II.

Table VI shows the means and confidence intervals of the adjustment vectors of the focus function extracted from on-, post-, and final-focus regions. These adjustment parameters are expressed as differences from the parameter values in the prefocus region shown in Table V. Both m and b in the on-focus region are magnified so that b is higher for the H tone but lower for the L tone while the absolute value of m is increased for both R and F tones. In contrast to the on-focus region, the parameter values in the postfocus region are compressed and lowered. In the final-focus region, the b adjustments are very small for the H tone but slightly larger for the L tone. Interestingly, there are no significant changes in on-focus λ adjustments as can be seen from their confidence intervals. This indicates that when the syllable is focused the approximation rate remains largely constant.

2. English

For English, parametric and adjustment vectors are obtained by averaging the individual vectors according to lexical stress, position in sentence, and focus. The positional differentiation is applied only in the prefocus region, which is to indirectly model the combined effect of new topic and sentence modality. Based on the findings of Xu and Xu (2005), we assume static targets for every syllable, except for

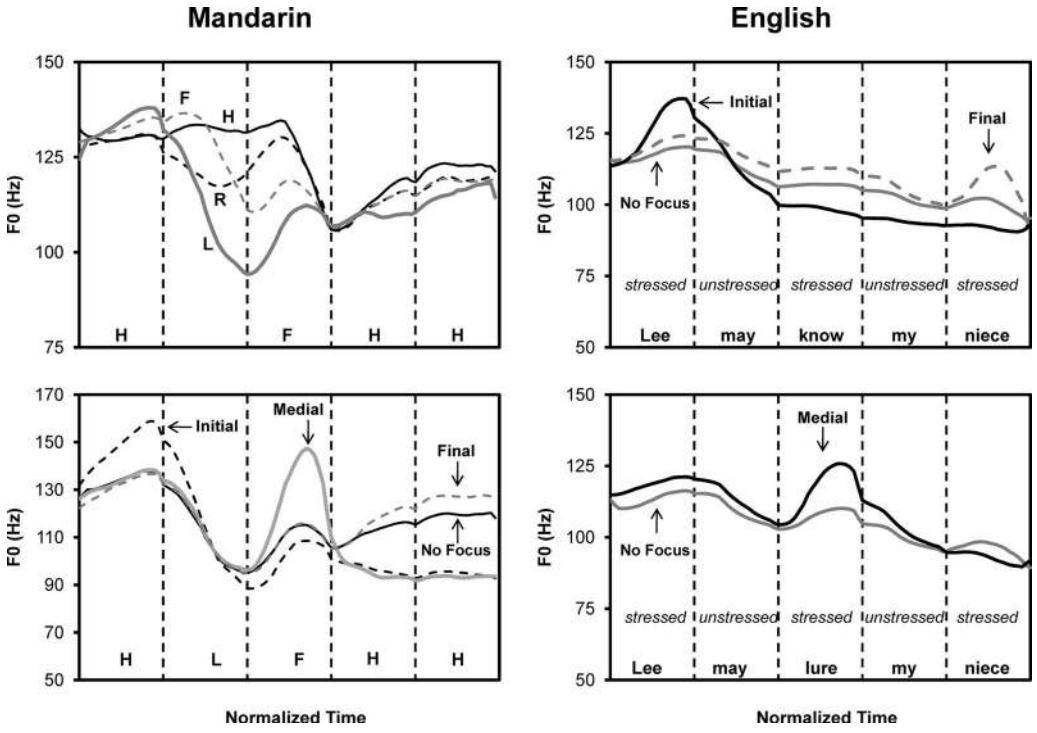

FIG. 10. Examples of naturally produced F_0 contours of Mandarin (left) and English (right) with tonal and focal variations. The vertical dashed lines mark the syllable boundaries. Adapted from Xu (1999) and Xu and Xu (2005), respectively.

TABLE VI. Means and confidence intervals of focus adjustment vectors of on-focus, postfocus, and final-focus regions obtained from the Mandarin dataset. These focus adjustment vectors are relative to the parametric vector of the tone function in Table V.

Focus location	Tone	Δm (st/s)	Δb (st)	$\Delta\lambda$ (s^{-1})
On focus	H	0	2.3 ± 1.1	-1.6 ± 3.5
	R	11.8 ± 3.6	0.6 ± 1.0	-3.9 ± 3.7
	L	0	-2.4 ± 1.8	1.0 ± 6.7
	F	-6.7 ± 2.5	1.2 ± 1.8	-1.2 ± 3.2
Postfocus	H	0	-5.6 ± 1.0	-11.2 ± 4.2
	R	-7.3 ± 2.9	-4.1 ± 0.7	7.8 ± 7.0
	L	0	-4.1 ± 1.4	-3.2 ± 5.4
	F	4.5 ± 3.3	-2.8 ± 1.5	2.5 ± 5.3
Final focus	H	0	-0.2 ± 0.8	-16.0 ± 2.5
	L	0	-2.1 ± 1.5	-4.6 ± 5.4

the sentence-final monosyllabic words and the focused word-final stressed syllables because their pitch targets in declarative sentences are likely to be [fall]. Table VII shows the means and confidence intervals of parametric vectors of the stress function in the prefocus region for each syllable position and stress condition, e.g., stress (stressed,1). The integer in the second argument of this stress function indicates syllable position. In Table VII, the b values are always higher for stressed syllables than for unstressed syllables even though focus adjustments are not yet involved in these prefocus regions. This is consistent with the finding that the effect of lexical stress on F_0 in English, though small in magnitude, is independent of focus (Xu and Xu, 2005) and audible to native listeners (Fry, 1958).

Table VIII shows the means and confidence intervals of the focus adjustment vectors obtained from on-, post-, and final-focus regions. While the values of postfocus adjustment b are all negative, the on-focus and final-focus adjustments are positive for stressed syllables but negative for unstressed syllables. This is consistent with the findings of Xu and Xu (2005). It is interesting to note that the on-focus $\Delta\lambda$ values are mostly negative. It could be partially a correction of the excessively large values for prefocus syllables seen in Table VII as compared to those in Table V for Mandarin. Speculatively, when the actual F_0 excursion size is very small, the estimation of λ tends to be larger than reality. This should be examined in future research. Also, the postfocus $\Delta\lambda$ values are positive for an unstressed syllable and no change for a stressed syllable, which differs from the reduction in postfocus word strength in the Stem-ML found by Shih and Kochanski (2003). This is also an issue worth exploring in future research.

TABLE VII. Means and confidence intervals of parametric vectors of the stress function obtained from the English dataset.

Syllable position	Stress	m (st/s)	b (st)	λ (s^{-1})
1	Unstressed	0	0.1 ± 0.8	72.8 ± 6.7
	Stressed	0	1.3 ± 0.9	48.1 ± 9.1
2	Unstressed	0	-1.2 ± 0.9	41.0 ± 5.9
3	Unstressed	0	-2.0 ± 0.7	51.4 ± 7.2
	Stressed	0	-1.2 ± 0.6	62.6 ± 11.8
4	Unstressed	0	-2.8 ± 0.6	43.3 ± 6.7
5	Unstressed	0	-5.3 ± 1.2	58.3 ± 9.9
	Stressed	0	-1.3 ± 1.2	49.8 ± 12.0

TABLE VIII. Means and confidence intervals of adjustment vectors of the focus function obtained from the English dataset. They are derived relative to the parametric vectors in Table VII.

Focus location	Stress	Δm (st/s)	Δb (st)	$\Delta\lambda$ (s^{-1})
On focus	Unstressed	0	-1.1 ± 0.7	-5.2 ± 3.7
	Stressed	0	2.9 ± 1.1	-10.6 ± 2.4
Postfocus	Unstressed	0	-1.9 ± 0.9	14.3 ± 5.0
	Stressed	0	-2.9 ± 1.0	1.4 ± 11.1
Final focus	Unstressed	0	-1.7 ± 2.7	0.9 ± 13.5
	Stressed	0	2.3 ± 1.6	-24.0 ± 8.7

Table IX shows the parametric vectors of the stress and focus functions for the sentence-final monosyllabic words and the focused word-final stressed syllable. Here the values of m are always negative regardless of focus condition, although the search space was not limited to below zero, indicating steep falling targets (Pierrehumbert, 1980; Xu and Xu, 2005). The postfocus adjustment of m is positive, indicating a decrease in the slope due to pitch range reduction.

Overall, the parameters obtained for both Mandarin and English are consistent with the results of systematic acoustic analyses for the respective corpora (Xu, 1999; Xu and Xu, 2005). This is initial indication that the analysis-by-synthesis method employed in the present study is effective.

C. Model evaluation by assessing synthesis quality

The effectiveness of the qTA model was further evaluated in two ways: (a) numerical assessment of closeness of fit between synthesized and natural F_0 and (b) perceptual identification of tone and focus as well as judgment of naturalness by native speakers of Mandarin and English.

1. Numerical assessment

The tests were conducted using a leave-one-out cross-validation scheme. This is to assure the reliability of the evaluation by testing the robustness of the qTA model against interspeaker variability. Each time, the data of one

TABLE IX. Means and confidence intervals of parametric vectors of the stress function and the adjustment vectors of the focus function for the exceptional cases where the pitch target is dynamic. These parametric vectors are also derived from the English corpus. The symbol Δ indicates that the parameters in that row are relative to the prefocus parameters.

Stress	Focus	Δm (st/s)	Δb (st)	$\Delta\lambda$ (s^{-1})
Focused word-final stressed syllable	On focus (Δ)	-81.1 ± 14.8	3.5 ± 1.4	-20.5 ± 1.3
Sentence-final monosyllabic word	Prefocus	-90.2 ± 11.8	-4.3 ± 1.5	38.4 ± 7.9
	Postfocus (Δ)	27.1 ± 14.6	-1.0 ± 1.9	-4.6 ± 5.6
	Final focus (Δ)	-22.8 ± 9.7	0.3 ± 1.6	-11.6 ± 2.7

TABLE X. Average RMSE and correlation coefficients in the Mandarin simulations. Synthesized F_0 used in these comparisons are those generated by resynthesis, tone function, focus function, and positional effect.

Imposed function	No. of parametric vector	RMSE (st)	Correlation
Resynthesis	19 200	0.56	0.92
Tone	4	2.84	0.74
Tone+position	20	2.46	0.75
Tone+focus	16	2.24	0.77
Tone+focus+position	80	2.16	0.78

speaker were circularly selected as the test set while those of the rest of the speakers formed the training set. Using this cross-validation scheme, the experiment repeated eight times, thus maximizing the chances of detecting the worst errors. Two measurements were used to assess the closeness of fit between synthesized and natural F_0, RMSE and Pearson's correlation coefficient. RMSE measures the difference between natural and synthesized F_0 contours while correlation coefficient indicates the linear relationship between them. It should be noted that the correlation coefficients were used for evaluating F_0 contours, not the model parameters. This evaluation matrix is the same as those reported in previous modeling attempts (Taylor, 2000; Dusterhoff *et al.*, 1999; Jilka *et al.*, 1999). Positive high correlation indicates consistency between original and synthesized F_0 contours not only in height but also in contour shapes. The semitone rather than hertz scale is used for measuring RMSE so that the results from different speakers can be assessed together. In contrast, correlation coefficients are computed in hertz to maximize the F_0 discrepancy. The training phase began with automatically extracting the parametric vectors from each utterance. The resulting parametric vectors were then summarized according to the functions to be tested. In the testing phase, RMSE and correlation coefficients for each sentence were obtained for each function.

The natural F_0 contours were first compared to the resynthesized contours, i.e., those generated with the parameters extracted from each individual sentence itself. Then, F_0 contours of two synthesized communicative functions were compared to the natural F_0: tone (Mandarin) or stress (English) and focus. Also, two positional dependencies were tested: syllable position (F_0 generated with parameters averaged for each position in a sentence) and tonal context (F_0 generated with parameters averaged for each tonal context). The syllable position dependency was tested for both Mandarin and English. The tonal context dependency was tested only for Mandarin.

2. Perceptual evaluation

For the perceptual evaluation experiments, the testing datasets were generated based on the communicative functions simulated in the present study. To test the effectiveness of duration modification in the focused syllable, the duration adjustment of the focus function was also included in the perceptual test. For the Mandarin tests, the following stimuli were generated.

Tone:	mao (H) mi (H/R/L/F) mo (H/R/F) ma (L) dao (H)	$4\times3=12$.
Focus:	neutral/initial/medial/final	$12\times4=48$.
Synthesis:	original/ synthetic without focus duration/ synthetic with focus duration	$48\times3=144$

The natural speech material was recorded specifically for the present study by the second author, who was one of the speakers in Xu (1999). In this recording, unlike in Xu (1999), the third syllable was always /mo/ regardless of its tone so as to guarantee minimal tonal contrasts for reliable assessment of tone identification at this sentence location.

For the Mandarin perceptual test, qTA parameters for tone and focus were extracted from each utterance. For each stimulus sentence, F_0 was synthesized and then used to replace the F_0 of the host utterance. There were 12 host utterances, in which the tones of the second and third syllables were varied. The focus-specific F_0 contours generated for each focus condition were used to replace the original contours of these utterances using pitch-synchronous overlap and add (PSOLA) method in PRAAT (Boersma, 2001). The importance of focus duration is tested by modifying the syllable duration of the segmental data of the original speech. Nine native Mandarin listeners participated in the test. The test was set up as a web-based program. The setup of the test consisted of two main sections. In the first section, listeners were instructed to identify the tones of the second and third syllables by selecting the Chinese character with the specific tone. In the second section, they were instructed to identify the focused word (i.e., the word being emphasized) and to judge whether the utterance was naturally spoken or synthesized. For each question in the test, the natural and synthesized sounds were randomly presented to the listeners. Listeners could listen to the sample sounds as many times as they preferred.

The conditions in the English perception experiment were focus, sentence, and synthesis method. There were eight natural speech utterances, including four neutral focus utterances from four different sentences, two initial focus utterances (S1 and S2), one medial focus utterance (S3), and one final focus utterance (S4). For speech with synthesized F_0, there were 32 test utterances according to the following composition.

Focus:	neutral/initial/medial/final	$4=4$.
Sentence:	[four different sentences] S1/ S2/S3/S4	$4\times4=16$.
Duration:	with focus duration/without focus duration	$16\times2=32$.

The natural speech materials used in making the stimuli were drawn from the English dataset used in the present study. Eight natural utterances were selected from the speaker with lowest resynthesis RMSE in the numerical assessment. Thus, there were totally 40 test utterances. To synthesize the test stimuli, we used only four neutral focus ut-

TABLE XI. Average RMSE and correlation coefficients for adding context dependency in the simulation of the Mandarin dataset.

Imposed function	No. of parametric vector	RMSE (st)	Correlation
Tone	4	2.84	0.74
Tone+preceding context	16	2.57	0.76
Tone+following context	16	2.50	0.77
Tone+focus	16	2.24	0.77
Tone+focus+preceding context	68	2.17	0.78
Tone+focus+following context	68	2.21	0.77

terances from four different sentences and modified the syllable duration of the segmental data of the original speech using PRAAT. Fourteen native English listeners participated in this listening experiment: ten of them were American English speakers and four were British English speakers. During the test, participants were asked to identify the focused words and to judge whether the test utterances were naturally spoken or synthesized. The test was set up in the same way as the Mandarin perception test.

3. Evaluation results

a. Mandarin Table X shows RMSE and correlation coefficients of the resynthesis and synthesis with successively added function-specific components. The results of resynthesis of the same utterances show very low error rates and very high correlations, indicating that the qTA-regenerated F_0 contours can fit the original contours quite well. From resynthesis to synthesis with function-specific parameters there is a general increase in error rate, but note also the dramatic reduction in the number of parameter vectors used (from 19 200 to 4–80). The tone-only condition has the highest error rate and lowest correlation. There is a slight improvement when either focus or positional specification was added. It is obvious that focus is the more effective constraint, as the result of adding focus specification is better than adding position specification. Moreover, the results of adding both focus and position are not significantly different in terms of RMSE from those of adding focus only [$F(1,15)=0.03$, $p=0.858$]. The more concrete evidence is in the analysis of correlation, which indicates that only the focus function significantly affects the correlation [$F(1,31)=5.21$, $p=0.030$]. This is probably because the focus specifications have already implicitly included the positional effect when calculating the adjustment vectors for different focus regions.

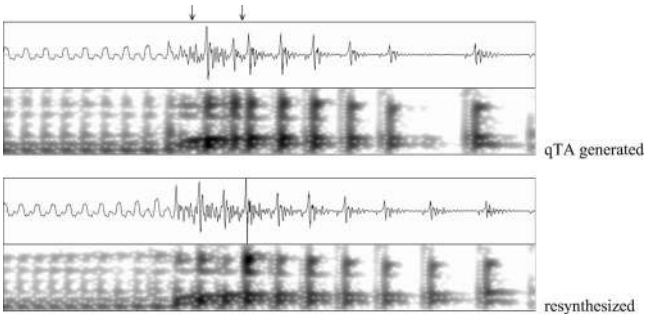

FIG. 12. Upper panel: Aperiodicity in a synthetic sentence. The arrows point to the two locations where the periods are either exceptionally long or exceptionally short. Lower panel: The resynthesized original sentence, where no strong aperiodicity is seen in the same locations.

Table XI shows the results of adding tonal context dependency to the model. The improvements are not significantly different between the preceding and following tone contexts, although they both show a slight but insignificant reduction in error rate. There are no further improvements in adding tonal context whether or not focus has been added [RMSE:$F(2,47)=0.23$, $p=0.795$; correlation: $F(2,47)=1.39$, $p=0.259$]. The focus function, however, significantly affects the correlation [$F(1,47)=4.65$, $p=0.036$]. This also indicates the importance of the focus function in determining the F_0 contours.

Figure 11 shows the results of the perceptual evaluation of natural and synthetic F_0 in Mandarin speech as described earlier. Listeners could identify the tones equally well in both syllable positions [$F(1,35)=0.54$, $p=0.470$]. There is no significant difference in identification rate between natural and synthesized F_0 [$F(1,35)=1.05$, $p=0.317$]. Moreover, there was no significant difference in focus identification between natural and synthesized F_0 whether or not natural duration adjustment was applied to the synthetic speech [$F(2,26)=2.36$, $p=0.127$]. However, naturalness perception differed between natural and synthesized F_0 [$F(2,26)=6.07$, $p=0.011$]. A close examination of the synthetic sentences that had low naturalness scores found cases of acoustic discontinuity. An example is shown in Fig. 12, where the locations of aperiodicity in the synthetic speech are indicated by the arrows. Such aperiodicity is not seen in

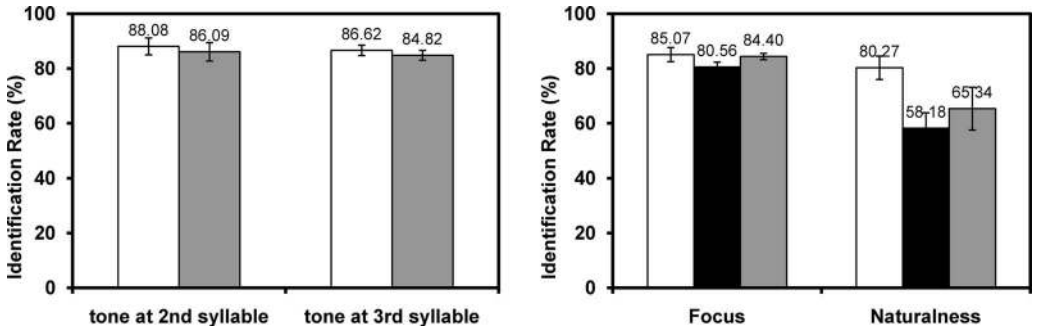

FIG. 11. Means (bars and the numbers above them) and standard errors (vertical lines) of identification rates in the Mandarin perceptual evaluation. The left graph shows averaged tone identification results while the right graph shows results of focus identification and naturalness evaluation. In the left graph, the white and gray bars indicate rate of tone identification for natural and synthetic F_0. In the right graph, the white, black, and gray bars indicate the results of focus identification and naturalness rating for natural F_0, synthetic F_0 without focus duration, and synthetic F_0 with focus duration, respectively.

TABLE XII. Averaged RMSE and correlation coefficients in the simulation of English dataset.

Imposed function	No. of parametric vector	RMSE (st)	Correlation
Resynthesis	14 224	0.32	0.83
Stress	4	1.93	0.75
Stress+position	11	1.71	0.78
Stress+focus	12	1.68	0.77
Stress+focus+position	18	1.57	0.78

the resynthesized original shown in the lower panel. This difference in synthesis quality is probably due to the ineffectiveness of the PSOLA algorithm in modifying pitch and duration at the same time.

b. English Table XII shows the results of numerical assessment for the English synthesis. For the resynthesized F_0 the errors are very low while the correlation coefficients are very high. Simulating only the stress function leads to higher errors and slightly lower correlations. This is similar to the case of the tone-only simulation in Mandarin in Table X. Similar to the Mandarin tests, there is no significant improvement by adding positional dependency when the focus function has already been used [RMSE: $F(1,15)=1.07$, $p=0.318$; correlation: $F(1,15)=0.34$, $p=0.571$]. There is, however, a significant main effect of adding positional dependency, but only in terms of correlation [$F(1,31)=4.94$, $p=0.034$]. An overall F_0 declination related to a combined effect of new topic and sentence modality, as discussed earlier, is the plausible underlying mechanism of this improvement.

Figure 13 shows the perception results of focus identification and naturalness evaluation of the natural and synthetic F_0 for the English dataset. There is a significant difference in focus identification between natural and synthetic F_0 [$F(2,41)=7.02$, $p=0.004$]. Listeners could identify focus significantly better from natural than from synthetic F_0 without duration adjustment [$F(1,27)=25.49$, $p<0.001$]. However, they could identify focus equally well from natural and synthetic F_0 with duration modification [$F(1,27)=0.62$, $p=0.444$]. Listeners also perceived a difference between synthesis methods [$F(2,41)=13.10$, $p<0.001$], but they did not recognize the difference in natural and synthetic F_0 with focus duration [$F(1,27)=1.82$, $p=0.200$].

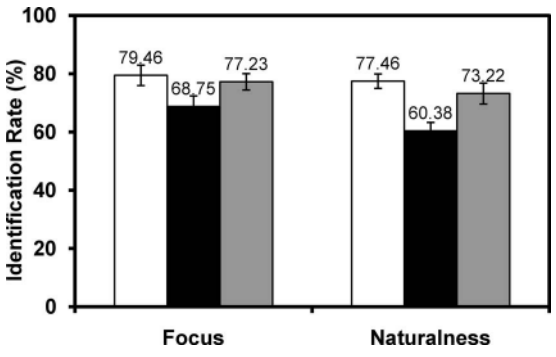

FIG. 13. Means (bars and the numbers above them) and standard errors (vertical lines) of focus identification rate and naturalness evaluation in the English perception tests. White, black, and gray bars correspond to sentences that are natural, synthetic without focus duration, and synthetic with focus duration, respectively. The vertical line in each bar indicates standard error.

V. DISCUSSION

The adequacy of any theoretical understanding of a natural phenomenon can be best measured in terms of the amount of details it can predict. This should also be true of theories about tone and intonation. Of the various tonal and intonational theories proposed so far (e.g., Bolinger, 1986; O'Connor and Arnold, 1961; Ladd, 1996; Pierrehumbert, 1980; Pike, 1945), few, if any, have been specific enough to allow full numerical testing of their adequacy. Although the autosegmental-metrical model has been used in synthesizing intonation as reported by Pierrehumbert (1981), no numerical tests were conducted to evaluate the quality of the synthesis. Meanwhile, there have also been many quantitative models of tone and intonation, as discussed in Sec. I, but none has been designed to directly test existing theories. While it is true that quantitative modeling can offer only an approximation to reality, theories without sufficient quantitative specifications can at best provide even coarser approximations to reality. The major goal of the present study is to quantify the theory about tone and intonation embodied in the PENTA model, which hitherto has not been fully numerical, and to subject it, through this quantification effort, to more rigorous testing than has been done before.

The PENTA model assumes that tone and intonation serve to convey communicative functions through specific encoding schemes that are implemented by the articulatory system, as sketched in Fig. 2. The model thus consists of two core components: an articulatory mechanism and a set of communicative functions. The articulatory mechanism assumed in PENTA is syllable-synchronized sequential TA, as depicted in Fig. 1. With this mechanism, the articulatory system asymptotically approaches the underlying targets that are either static or dynamic. The approximation is always synchronized with the syllable. But due to inertia, articulatory states are transferred onward across syllable boundaries. Although quite specific, this conceptualization needs to be quantified into a dynamic system that is biophysically plausible. The communicative functions are considered in PENTA as the driving force of the system. They are assumed to be parallel to each other, each with a unique encoding scheme in terms of one or more of the TA parameters. The uniqueness and complexity of these functions entail that each of them can be properly simulated only with sufficient knowledge about its temporal domain of operation, the TA parameters involved, and the value range of the parameters, e.g., whether the target is static or dynamic.

To develop a biomechanically plausible dynamic system, the simulation is best done at a level where the link between articulatory control and communicative functions is the most direct, although many levels of biophysical processes are apparently involved. Simulation at the level of individual muscular forces, as done in the CR model (Fujisaki *et al.*, 2005), would entail variable parameter values depending on the tonal contexts, as the individual mus-

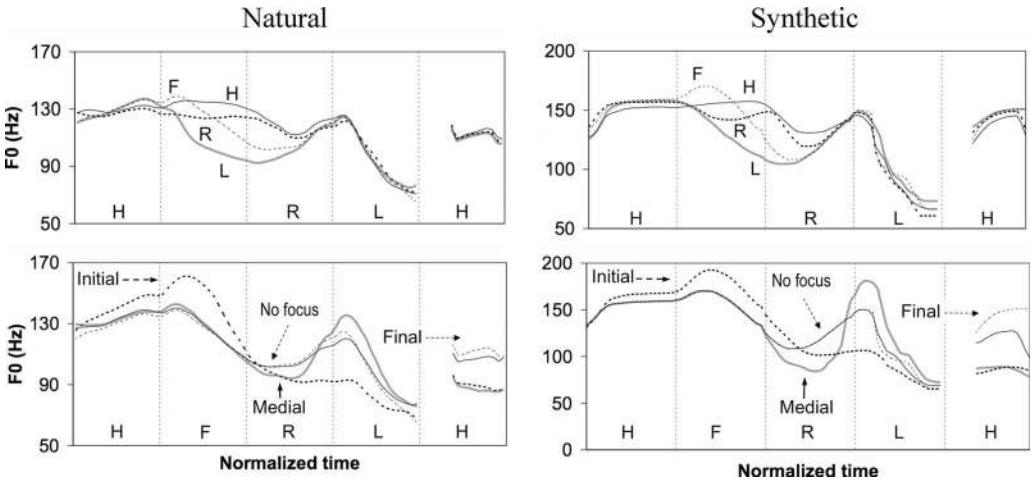

FIG. 14. Examples of natural and synthetic F_0 contours of tone and intonation in Mandarin. In the top plots, the sentences differ in the tone of the second syllable, with focus always on the last two syllables. In the bottom plots, the sentences differ in focus location: the first (disyllabic) word, the second (monosyllabic) word, the final (disyllabic) word, or none of the words.

cular forces have to be differentially adjusted according to the distance to be covered between the initial and targeted articulatory states. Simulation at a level where F_0 is controlled without the mediation of an articulatory mechanism, as done in the tilt model (Taylor, 2000) and SFC model (Bailly and Holm, 2005), would miss critical dynamic details due to articulatory constraints, such as unidirectional left-to-right assimilatory influence (Gandour *et al.*, 1994; Wong, 2006; Xu, 1997, 1999) or peak delay across syllable boundaries (Xu, 1998, 2001), both due to the physical constraint of inertia. The qTA model developed in the present study implements a third-order critically damped system, which generates F_0 contours through syllable-synchronized sequential TA. This strategy severely constrains the degrees of freedom of the model, limiting its variable control parameters to only three, each corresponding directly to a TA parameter in PENTA. Automatic parameter extraction through analysis by synthesis in the present study has found these parameters to remain largely invariant across different tonal contexts, especially the preceding tonal context which is known to cause a large amount of contextual variability (Gandour *et al.*, 1994; Wong, 2006; Xu, 1997, 1999). This can be seen in Fig. 14 (upper row). Also can be seen in Fig. 14 is that the phenomenon of peak delay (Xu, 2001) in a RL sequence is effectively simulated without using peak location as a control parameter.

The correspondence of the qTA parameters to the TA parameters in the PENTA model means that the assumptions of the PENTA model about the interorthogonality between the communicative functions and about the uniqueness of the function-specific encoding schemes can be tested through supervised learning. To that end, we tested the model with two natural speech corpora (Xu, 1999; Xu and Xu, 2005). The parameter extraction was guided by knowledge obtained from previous production studies (for English: Cooper *et al.*, 1985; Xu and Xu, 2005; for Mandarin: Liu and Xu, 2005; Xu, 1997, 1999), which was used to limit the value range of the parameters. For example, based on previous research on focus, separate parameter sets were obtained for prefocus, on-focus, postfocus, and final-focus regions. Note that restricting the parameters' range this way during training is actually risky because anything faulty in the assumptions behind the restrictions would increase the chance of generating large errors during testing. The results of the testing were nevertheless encouraging. The numerical evaluations showed that even when applied to speakers not included in the training, the error rates were comparable with those of previous studies (Fujisaki *et al.*, 2005; Hirst and Espresser, 1993; Kochanski and Shih, 2003; Pierrehumbert, 1981; Taylor, 2000). More significantly, the results of the perceptual evaluations show that not only can native listeners identify the functional categories from model-based synthesis just as well as from natural speech, but also they could not reliably distinguish utterances with synthetic F_0 from those with natural F_0 in terms of naturalness, especially in the case of English. These results not only demonstrate the effectiveness of the qTA model but also provide support for the validity of the theoretical assumption of the PENTA model that tone and intonation are in essence articulatorily encoded communicative functions.

A secondary goal of the present study is to develop an intonation generation system that is applicable in speech technology. There have been a number of attempts to apply automatic trainings to articulatory-oriented parameters in intonation synthesis (Fujisaki *et al.*, 2005; Kochanski and Shih, 2003; Mixdorff, 2000). Although they can approximate F_0 contours quite accurately, the lack of complete framework to implement the communicative functions in these models makes it difficult to effectively integrate them in speech processing systems. The simple form of underlying pitch targets consisting of only two parameters in qTA also makes the process of TA more straightforward than those that are more complex, e.g., templates and commands. Although the practicality of analysis by synthesis is limited due to a large search space, the articulatory constraints imposed on the automatic parameter extraction process effectively reduce the search space, as shown in this study. Thus for speech technology, the application of qTA in speech synthesis may improve the intonation components of the synthesizers.

Despite the overall quality of the simulation in the present study, a number of caveats need to be mentioned. The first is that the strategy used here is to simulate only functions that have been systematically studied in empirical research, and to do so by imposing specific restrictions on the search space for the model parameters based on the empirical findings. The simulation process is therefore not fully "automatic." It is our understanding, however, that our current knowledge level does not yet allow us to build a fully automatic process without sacrificing the quality of simulation. This is because there are many communicative functions that are encoded by intonation as discussed in the Introduction, and without knowing full well what they are and which of them have occurred in a particular utterance, it is difficult even for trained human labelers to label them consistently (Wightman, 2002; Xu, 2006). On the other hand, once a given specific communicative function has been recognized, it is just one step further from the current strategy of imposing *a priori* limits on the search space to one that is able to discover the TA parameters used by the function and the range of their values. This could be done by taking advantages of recent development in unsupervised learning of phonetic categories (Gauthier *et al.*, 2007; Guenther, 1994). It would be even more desirable for future development of the model to attain the ability to automatically discover the communicative functions encoded in a particular language. As far as we can see, this could be done only if we have accumulated a substantial amount of knowledge through further research.

Second, several phenomena reported for tone and intonation that are likely due to articulatory mechanisms have not yet been simulated in the current version of the qTA model. These include anticipatory dissimilation, i.e., the raising of F_0 by a following L tone (Gandour *et al.*, 1994; Laniran and Clements, 2003; Wong, 2006; Xu, 1997, 1999), post-L bouncing (Chen and Xu, 2006; Pierrehumbert, 1980), whereby F_0 is raised by a preceding L tone, and dynamic delay, whereby the initial approximation of a dynamic pitch target is delayed at a slow speech rate so as to guarantee that the slope of the target is not reached earlier than the end of the syllable (Wong, 2006; Xu, 1998, 2001). Future development of the qTA model will need to find ways to simulate these mechanisms. Because each of these mechanisms seems to affect only a particular aspect of the TA process (Xu, 2006), they will likely to be modeled as additional mechanisms added on top of the core algorithm of qTA.

Third, one of the TA parameters, namely, duration, was only partially tested in the present study. It was tested in the perceptual evaluations by comparing focus-specific syllable duration with non-focus-specific duration. The former was found to generate F_0 contours that better match natural F_0 in conveying focus. While this is an encouraging result, what is lacking in general is a theoretical link between duration and specific communicative functions, for which existing research so far has provided only a vague picture (see Shattuck-Hufnagel and Turk, 1996 for a review). Further studies aimed at revealing that the functional sources of duration patterns are needed [see Xu, (2008) for further discussion].

Finally, only a limited number of communicative functions have been modeled in the current project. Many more that have been discussed in literature (Bolinger, 1989; Hirschberg, 2002; Hirst, 2005; Kohler, 2005; Xu, 2005, 2008) such as sentence type (e.g., statement versus question), topic and turn taking, speaking style, emotion, to name just a few, have not been tested. As discussed in the Introduction, qTA, as a quantitative implementation of PENTA, was developed exactly for the purpose of modeling multiple communicative functions with a limited number of articulatory-based parameters. The present results seem to suggest that it will provide a promising tool for modeling many more communicative functions in future research.

VI. CONCLUSION

We have proposed in this paper a qTA model which implements the theoretical PENTA model and simulates F_0 contours as the process of TA. The model simulates surface F_0 by adjusting parameters of local TA, including the height and slope of the pitch targets and the rate of approaching individual targets. There are only three free parameters in qTA, and their controls are all directly linked to specific communicative functions such as lexical tone, lexical stress, focus, sentence modality, and new topic, although only the first three were explicitly modeled in the present study. We tested the model by training it with Mandarin and English speech data using an automated analysis-by-synthesis procedure. The F_0 simulations were evaluated both numerically and perceptually. The accuracy of resynthesis by the model was high, with RMSE of 0.56 st and correlation of 0.92 for Mandarin, and RMSE of 0.32 st and correlation of 0.83 for English. The quality of synthesizing specific communicative functions, including tone in Mandarin, lexical stress in English, and focus in both languages, was comparable to previous studies even when the trained parameters were applied to speakers not included during training, with RMSE of 2.24 st and correlation of 0.77 for Mandarin, and RMSE of 1.68 st and correlation of 0.77 for English. More importantly, listeners' perceptual identification of tone, lexical stress, and focus, and their judgment of the naturalness of speech is nearly identical between natural and synthetic F_0. These results demonstrate the validity of the assumptions underlying qTA and suggest that it can be used as an effective tool both for theoretical study of tone and intonation and for generating F_0 contours in automatic speech synthetic systems.

[1]Arvaniti, A., and Ladd, D. R. (1995). "Tonal alignment and the representation of accentual targets," Proceedings of the 13th International Congress of Phonetic Sciences, Stockholm, Vol. **4**, pp. 220–223.
[2]Arvaniti, A., Ladd, D. R., and Mennen, I. (1998). "Stability of tonal alignment: The case of Greek prenuclear accents," J. Phonetics **26**, 3–25.
[3]Atterer, M., and Ladd, D. R. (2004). "On the phonetics and phonology of "segmental anchoring" of F_0: Evidence from German," J. Phonetics **32**, 177–197.
[4]Bailly, G., and Holm B. (2005). "SFC: A trainable prosodic model," Speech Commun. **46**, 348–364.
[5]Bernstein, N. (1967). *The Co-Ordination and Regulation of Movement* (Pergamon, London).
[6]Black, A. W., and Hunt, A. J. (1996). "Generating F_0 contours from ToBI labels using linear regression," Proceedings of ICSLP 1996, Philadelphia, PA, pp. 1385–1388.
[7]Boersma, P. (2001). "Praat, a system for doing phonetics by computer,"

Glot International **5**, 341–345.
[8]Bolinger, D. (1986). *Intonation and Its Parts: Melody in Spoken English* (Stanford University Press, Palo Alto, CA).
[9]Bolinger, D. (1989). *Intonation and Its Uses: Melody in Grammar and Discourse* (Stanford University Press, Stanford, Ca).
[10]Botinis, A., Bannert, R., and Tatham, M. (2000). "Contrastive tonal analysis of focus perception in Greek and Swedish,." in *Intonation: Analysis, Modelling and Technology*, edited by A. Botinis (Kluwer Academic, Boston), 97–116.
[11]Chen, A., Gussenhoven, C., and Rietveld, T. (2004). "Language-specificity in the perception of paralingistic intonational meaning," Iowa Dent. Bull. **47**, 311–349.
[12]Chen, Y., and Xu, Y. (2006). "Production of weak elements in speech: Evidence from F_0 patterns of neutral tone in standard Chinese," Phonetica **63**, 47–75.
[13]Cooper, W. E., Eady, S. J., and Mueller, P. R. (1985). "Acoustical aspects of contrastive stress in question-answer contexts," J. Acoust. Soc. Am. **77**, 2142–2156.
[14]d'Alessandro, C., and Mertens, P. (1995). "Automatic pitch contour stylization using a model of tonal perception," Comput. Speech Lang. **9**, 257–288.
[15]de Jong, K. (1994), "Initial tones and prominence in Seoul Korean," in *Working Papers in Linguistics* edited by S.-H Lee and S.-A. Jun (The Ohio State University Department of Linguistics), Vol. **43**.
[16]Dusterhoff, K. E., Black, A. W., and Taylor, P. (1999). "Using decision tree within the tilt intonation model to predict F_0 contours," Proceedings of EUROSPEECH'99, Budapest, pp. 1627–1630.
[17]Erickson, D. M. (1976). "A physiological analysis of the tones of Thai," Ph.D. thesis, University of Connecticut.
[18]Erickson, D., Honda, K., Hirai, H., and Beckman, M. E. (1995). "The production of low tones in English intonation," J. Phonetics **23**, 179–188.
[19]Fry, D. B. (1958). "Experiments in the perception of stress," Leeds Dent. J. **1**, 126–152.
[20]Fujimura, O. (2000). "Rhythmic organization and signal characteristics of speech," Proceedings of ICSLP 2000, Beijing, Vol. **1**, pp. 29–35.
[21]Fujisaki, H. (1974). "Formulation of the coarticulatory process in the formant frequency domain and its application to automatic recognition of connected vowels," Proceedings of the Speech Communication Seminar, Almqvist & Wiksell, Uppsala, Stockhom, pp. 385–392.
[22]Fujisaki, H. (1983). "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*, edited by P. F. MacNeilage (Springer-Verlag, New York), pp. 39–55.
[23]Fujisaki, H. (2003). "Prosody, information, and modeling: With Emphasis on Tonal Features of Speech," Proceedings of Workshop on Spoken Language Processing, Mumbai, pp. 5–14.
[24]Fujisaki, H., Wang, C., Ohno, S., and Gu, W. (2005). "Analysis and synthesis of fundamental frequency contours of standard Chinese using the command-response model," Speech Commun. **47**, 59–70.
[25]Gandour, J., Potisuk, S., and Dechongkit, S. (1994). "Tonal coarticulation in Thai," J. Phonetics **22**, 477–492.
[26]Gauthier, B., Shi, R., and Xu, Y. (2007). "Simulating the acquisition of lexical tones from continuous dynamic input," J. Acoust. Soc. Am. **121**, EL190–EL195.
[27]Gribble, P. L., Mullin, L. L., Cothros, N., and Mattar, A. (2003). "Role of cocontraction in arm movement accuracy," J. Neurophysiol. **89**, 2396–2405.
[28]Gribble, P. L., and Scott, S. H. (2002). "Overlap of internal models in motor cortex for mechanical loads during reaching," Nature (London) **417**, 938–941.
[29]Gu, W., Hirose, K., and Fujisaki, H. (2007). "Analysis of tones in Cantonese speech based on the command-response model," Phonetica **64**, 29–62.
[30]Guenther, F. H. (1994). "A neural network model of speech acquisition and motor equivalent speech production," Biol. Cybern. **72**, 43–53.
[31]Gussenhoven, C. (2007). "Types of focus in English," in *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation*, edited by C. Lee, M. Gordon and D. Büring (Springer, New York), pp. 83–100.
[32]Hallé, P. A. (1994). "Evidence of tone-specific activity of the sternohyoid muscle in modern standard Chinese," Leeds Dent. J. **37**, 103–123.
[33]Hasegawa, Y., and Hata, K. (1992). "Fundamental frequency as an acoustic cue to accent perception," Leeds Dent. J. **35**, 87–98.
[34]Hirano, M. (1974). "Morphological structure of the vocal cord as a vibrator and its variations," Folia Phoniatr. **26**, 89–94.
[35]Hirschberg, J. (2002). "Communication and prosody: Functional aspects of prosody," Speech Commun. **36**, 31–43.
[36]Hirst, D. (2005). "Form and function in the representation of speech prosody," Speech Commun. **46**, 334–347.
[37]Hirst, D., and Espesser, R. (1993). "Automatic modelling of fundamental frequency using a quadratic spline function," Travaux de l'Institut de Phonétique d'Aix **15**, 75–85.
[38]Jilka, M., Möhler, G., and Dogil, G. (1999). "Rules for the generation of ToBI-based American English intonation," Speech Commun. **28**, 83–108.
[39]Kelso, J. A. S. (1982). "The process approach to understanding human motor behavior: an introduction," in *Human Motor Behavior: An Introduction*, edited by J. A. S. Kelso (Erlbaum, Hillsdale, NJ), pp. 3–19.
[40]Klatt, D. H. (1987). "Review of text-to-speech conversion for English," J. Acoust. Soc. Am. **82**, 737–792.
[41]Kochanski, G., and Shih, C. (2003). "Prosody modeling with soft templates," Speech Commun. **39**, 311–352.
[42]Kochanski, G., Shih, C., and Jing, H. (2003). "Quantitative measurement of prosodic strength in Mandarin," Speech Commun. **41**, 625–645.
[43]Kohler, K. (2005). "Timing and communicative functions of pitch contours," Phonetica **62**, 88–105.
[44]Krahmer, E., and Swerts, M. (2001). "On the alleged existence of contrastive accents," Speech Commun. **34**, 391–405.
[45]Laniran, Y. O., and Clements, G. N. (2003). "Downstep and high raising: Interacting factors in Yoruba tone production," J. Phonetics **31**, 203–250.
[46]Ladd, D. R. (1983). "Phonological features of intonational peaks," Language **59**, 721–759.
[47]Ladd, D. R. (1996). *Intonational Phonology* (Cambridge University Press, Cambridge).
[48]Lehiste, I. (1975). "The phonetic structure of paragraphs," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. E. G. Nooteboom (Springer-Verlag, New York), pp. 195–206.
[49]Lindblom, B. (1983) "Economy of speech gestures," in *The Production of Speech*, edited by P. MacNeilage (Springer, New York).
[50]Liu, F., and Xu, Y. (2005). "Parallel encoding of focus and interrogative meaning in Mandarin intonation," Phonetica **62**, 70–87.
[51]Mixdorff, H. (2000). "A novel approach to the fully automatic extraction of Fujisaki model parameters," Proceedings of ICASSP 2000, Istanbul, Vol. **3**, pp. 1281–1284.
[52]Mixdorff, H. (2004). "Quantitative tone and intonation modeling across languages," Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, pp. 137–142.
[53]Monsen, R. B., Engebretson, A. M., and Vemula, N. R. (1978). "Indirect assessment of the contribution of subglottal air pressure and vocal fold tension to changes in the fundamental frequency in English," J. Acoust. Soc. Am. **64**, 65–80.
[54]Ni, J., Kawai, H., and Hirose, K. (2006). "Constrained tone transformation technique for separation and combination of Mandarin tone and intonation," J. Acoust. Soc. Am. **119**, 1764–1782.
[55]Nolan, F. (2003). "Intonational equivalence: An experimental evaluation of pitch scales," Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, pp. 771–774.
[56]O'Connor, J. D., and Arnold, G. F. (1961). *Intonation of Colloquial English* (Longmans, London).
[57]Ohala, J. J. (1978). "Production of tone," in *Tone: A Linguistic Survey*, edited by V. A. Fromkin (Academic, New York), pp. 5–39.
[58]Palm, W. J., III (1999). *Modeling, Analysis, and Control of Dynamic Systems*, 2nd ed. (Wiley, New York).
[59]Pierrehumbert, J. (1980). "The phonology and phonetics of English intonation," Ph.D. thesis, MIT, Cambridge, MA.
[60]Pierrehumbert, J. (1981). "Synthesizing intonation," J. Acoust. Soc. Am. **70**, 985–995.
[61]Pierrehumbert, J., and Hirschberg, J. (1990). "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack (MIT Press, Cambridge, MA), pp. 271–311.
[62]Pierrehumbert, J., and Steele, S. (1990). "Categories of tonal alignment in English," Phonetica **47**, 181–196.
[63]Pike, K. (1945). *The Intonation of American English* (University of Michigan Press, Ann Harbor, MI).
[64]Prieto, P., van Santen, J., and Hirschberg, J. (1995). "Tonal alignment patterns in Spanish," J. Phonetics **23**, 429–451.
[65]Ross, K. N., and Ostendorf, M. (1999). "A dynamical system model for generating fundamental frequency for speech synthesis," IEEE Trans. Speech Audio Process. **7**, 295–309.
[66]Rump, H. H., and Collier, R. (1996). "Focus conditions and the promi-

nence of pitch-accented syllables," Leeds Dent. J. **39**, 1–17.
[67]Saltzman, E. L., and Kelso, J. A. S. (1987). "Skilled actions: A task dynamic approach," Psychol. Rev. **94**, 84–106.
[68]Saltzman, E. L., and Munhall, K. G. (1989). "A dynamical approach to gestural patterning in speech production," Ecological Psychol. **1**, 333–382.
[69]Shattuck-Hufnagel, S., and Turk, A. E. (1997). "A prosody tutorial for investigators of auditory sentence processing," J. Psycholinguist. Res. **25**, 193–247.
[70]Shih, C., and Kochanski, G. (2003). "Modeling intonation: Asking for confirmation in English," Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, pp. 551–554.
[71]Silverman, K., and Pierrehumbert, J. (1990). "The timing of prenuclear high accents in English," *Papers in Laboratory Phonology* (Cambridge University Press, Cambridge, UK), Vol. **I**, pp. 72–106.
[72]Sun, X. (2002). "The determination, analysis, and synthesis of fundamental frequency," Ph.D. thesis, Northwestern University.
[73]Swerts, M. (1997). "Prosodic features at discourse boundaries of different length," J. Acoust. Soc. Am. **101**, 514–521.
[74]'t Hart, J., Collier, R., and Cohen, A. (1990) *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody* (Cambridge University Press, Cambridge).
[75]Taylor, P. (2000). "Analysis and synthesis of intonation using the tilt model," J. Acoust. Soc. Am. **107**, 1697–1714.
[76]Titze, I. R. (1989). "On the relation between subglottal pressure and fundamental frequency in phonation," J. Acoust. Soc. Am. **85**, 901–906.
[77]van Heuven, V. J. (1994). "What is the smallest prosodic domain?," in *Papers in Laboratory Phonology*, edited by P. A. Keating (Cambridge University Press, Cambridge), Vol. **3**, pp. 76–98.
[78]van Santen, J., and Möbius, B. (2000). "A quantitative model of F_0 generation and alignment," in *Intonation: Analysis, Modelling and Technology*, edited by A. Botinis, (Kluwer, Dordrecht), pp. 269–288.
[79]Wang, B., and Xu, Y. (2006). "Prosodic encoding of topic and focus in Mandarin," Proceedings of Speech Prosody 2006, Dresden, pp. 313–316.
[80]Wightman, C. W. (2002). "ToBI or not ToBI," Proceedings of Speech Prosody 2002, Aix-en-Provence, pp. 25–29.
[81]Wong, Y. W. (2006). "Contextual tonal variations and pitch targets in Cantonese," Proceedings of Speech Prosody 2006, Dresden, pp. 317–320.
[82]Wong, Y. W., and Xu, Y. (2007). "Consonantal perturbation of F_0 contours of Cantonese tones," Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, pp. 1293–1296.
[83]Xu, Y. (1997). "Contextual tonal variations in Mandarin," J. Phonetics **25**, 61–83.
[84]Xu, Y. (1998). "Consistency of tone-syllable alignment across different syllable structures and speaking rates," Phonetica **55**, 179–203.
[85]Xu, Y. (1999). "Effects of tone and focus on the formation and alignment of F_0 contours," J. Phonetics **27**, 55–105.
[86]Xu, Y. (2001). "Fundamental frequency peak delay in Mandarin," Phonetica **58**, 26–52.
[87]Xu, Y. (2005). "Speech melody as articulatorily implemented communicative functions," Speech Commun. **46**, 220–251.
[88]Xu, Y. (2006). "Principles of tone research," Proceedings of Second International Symposium on Tonal Aspects of Languages, La Rochelle, pp. 3–13.
[89]Xu, Y. (2008). "Timing and coordination in tone and intonation: An articulatory-functional perspective," Lingua (2008).
[90]Xu, Y., and Sun, X. (2002). "Maximum speed of pitch change and how it may relate to speech," J. Acoust. Soc. Am. **111**, 1399–1413.
[91]Xu, Y., and Wallace, A. (2004) "Multiple effects of consonant manner of articulation and intonation type on F_0 in English," J. Acoust. Soc. Am. **115**, 2317.
[92]Xu, Y., and Wang, Q. E. (2001). "Pitch targets and their realization: Evidence from Mandarin Chinese," Speech Commun. **33**, 319–337.
[93]Xu, C. X., and Xu, Y. (2003). "Effects of consonant aspiration on Mandarin tones," J. Int. Phonetic Assoc. **33**, 165–181.
[94]Xu, Y., and Xu, C. X. (2005). "Phonetic realization of focus in English declarative intonation," J. Phonetics **33**, 159–197.
[95]Xu, C. X., Xu, Y., and Luo, L.-S. (1999). "A pitch target approximation model for F_0 contours in Mandarin," Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, pp. 2359–2362.
[96]Zemlin, W. R. (1988). *Speech and Hearing Science: Anatomy and Physiology* (Prentice-Hall, Englewood Cliffs, NJ).