Functional-oriented articulatory modeling of tones and intonations

Santitham Prom-on[1,2], Yi Xu[2,3] and Bundit Thipakorn[1]

[1]Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand
[2]Department of Phonetics and Linguistics, University College London, UK
[3]Haskins Laboratories, New Haven, CT, USA
santitham@cpe.kmutt.ac.th; yi@phon.ucl.ac.uk; bundit@cpe.kmutt.ac.th

Abstract

In this paper we report results of applying the quantitative target approximation model (qTA) [7] to simulate function-specific F_0 contours in Mandarin. The qTA model is based on a set of assumptions about the biophysical and neural control mechanisms of pitch production. To simulate F_0 contours for tone and focus, we extracted qTA parameters that are tone-specific and adjustment parameters that are focus-specific. The accuracy and effectiveness of this approach were tested through a series of synthesis experiments. In the baseline case, the results were fair with just tonal specifications. Further experiments showed additional improvements when the parameters became more functions-specific.

1. Introduction

Speech conveys communicative meanings through sounds generated by the human vocal apparatus. This means that effective prosodic modeling can be achieved by simultaneously simulating the articulatory process of F_0 production and the process of encoding communicative meanings. The articulatory process has been proposed as one of syllable-synchronized sequential target approximation (the TA model) [11], and the encoding process has been proposed as one of parallel control of the TA parameters by separate communicative functions (the PENTA model) [9]. To test the understanding represented by the two conceptual models, we developed the qTA model [7], a quantitative implementation of the TA model, whose parameters are suitable for encoding communicative functions. In this paper we report the results of our testing of the model. We will discuss, in particular, how the model may help to achieve three objectives that, we believe, are critical for any robust speech modeling: (a) simulating articulatory processes that are biophysically plausible, (b) automatic extraction of model parameters from natural speech, and (c) generating acoustic forms that convey specific communicative functions.

2. The qTA model for simulating biophysically plausible processes

The qTA model is the quantitative version of the Target Approximation (TA) model [11] which theoretically outlines the relationship between surface F_0 contours and underlying articulatory mechanisms. The qTA model is based on a number of biophysical assumptions that constrain the F_0 implementation process. In this section, we will summarize the major assumptions and the mathematical framework of the model. More discussions on the model assumption can be found in [7].

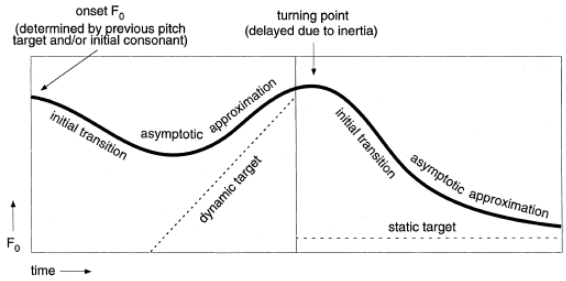

Figure 1: *An illustration of the theoretical pitch target approximation model proposed in* [9]

2.1. Assumptions

The qTA model assumes that F_0 is directly related to the tension of the vocal folds. This tension is controlled by two antagonistic muscle forces generated by the contraction of the cricothyroid, the thyroarytenoid and the strap muscles. Furthermore, since there is no observed oscillation in F0 movements, the tension control mechanism can be represented by an overdamped second order system.

As shown in Fig. 1, the TA model [11] describes F_0 trajectories as asymptotic movements toward successive targets. Because these are goal-oriented movements, they have to be controlled by a feedback mechanism to ensure that the movement reaches the desired goal. Hence, the qTA model is built as a time-delayed feedback control of an overdamped second-order system.

The goal of each F_0 movement takes the forms of either a static or a dynamic linear target as illustrated in Fig. 1. Each target can be represented by two parameters, slope and height. For example, in Mandarin, H, L, and N tones can be represented by static targets of different heights while R and L tones by dynamic targets with different slopes and heights.

Also depicted in Fig. 1, the implementation of each tone is assumed to be synchronous with the syllable [11]. This also leads to the further assumption that the state of articulation, in terms of F_0 level, velocity, and acceleration, is transferred from one syllable to the next, resulting in a smooth F_0 trajectory despite abrupt target shifts across syllable boundaries.

2.2. Mathematical Framework

Fig. 2 shows a block diagram of the qTA model. The target can be represented by a simple linear function.

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

where m and b denote the slope and height of the target, respectively. The behavior of the second-order system is

specified by two parameters, ζ and ω_n, namely, the damping ratio and the undamped natural frequency. ζ characterizes the responsiveness of the tension control. ω_n indicates the amount of effort used to implement the target. The double amplifier is used to compensate the halving reduction of the feedback so that the forced response of the system is the pitch target. The time-delayed feedback is approximated by the first-order Padé approximant which, as a result, increases the order of the overall system by one. Thus, the complete response of the model is in the third-order form:

$$F_0(t) = c_1 e^{r_1 t} + c_2 e^{r_2 t} + c_3 e^{r_3 t} + mt + b \quad (2)$$

where r_1, r_2, and r_3, are the roots of the homogeneous equation of the total differential equation. The coefficients c_1, c_2, and c_3 are solved from the initial conditions which include F_0 level, velocity, and acceleration.

Target, defined by Eq. (1), is the forcing function of the system, and is ideally reached at the end of the syllable. Given specific syllable duration and the amount of effort represented by ω_n, however, the target may not be reached.

Finally, the sequential target approximation of the TA model implies another critical mechanism of F_0 realization by the qTA model. As mentioned earlier, the coefficients in Eq. (2) are calculated from the initial conditions. Except in the first syllable, these initial conditions are transferred from the final state of the previous syllable:

$$\begin{aligned} F_0(0)_i &= F_0(t_{i-1}^{final})_{i-1} \\ F_0'(0)_i &= F_0'(t_{i-1}^{final})_{i-1} \\ F_0''(0)_i &= F_0''(t_{i-1}^{final})_{i-1} \end{aligned} \quad (3)$$

where the conditions with subscription i denote the initial conditions of the i^{th} syllable of that sentence. This transfer mechanism reflects the propagation of the laryngeal state and its dynamics across the syllable boundary [9].

3. Automatic Parameter Extraction

Like what has been done previously [1-3], the parameter extraction was done with an automatic analysis-by-synthesis optimization algorithm. This algorithm varies the parameter values in the specified search space and obtains the parameter set with the lowest sum square error. Also, there is a need to specify appropriate search spaces, without which the optimization process can be easily stuck at a local minimum.

Thanks to the biophysical assumptions just discussed, we adopted principled ways to restrict the degrees of freedom of the system, which reduces the difficulty in parameter extraction. Unlike in previous efforts, various restriction are imposed on the optimization process. First, ζ is arbitrarily fixed at a constant value in an overdamping range ($\zeta > 1$), e.g. 1.5 in our tests. Consequently, the effort control parameter, ω_n, is mathematically limited by T_d, which is set to 5 ms. Although the time delay of the feedback loop may be too small for the auditory perception, it is possible that this small time delay comes from the forward model within the brain [5]. Second, following the target assumption, we specified the search space for m to depend on the tonal category: zero for H, L, and N, positive for R, and negative for F. Third, also based on the target assumption, we restricted the search space of b to center around the final F_0 level of each syllable with a small vertical range, because that is where surface F_0 gets closest to the target.

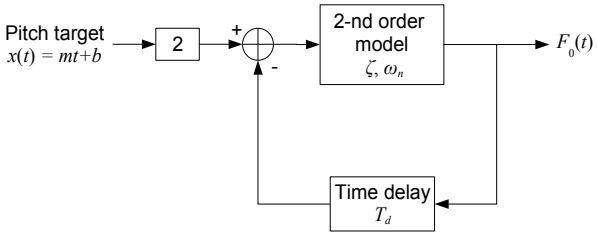

Figure 2: *A block diagram of the quantitative target approximation model proposed in* [7].

4. Deriving function-specific parameters

The PENTA model [9] assumes that speech prosody aims to convey communicative functions encoded in parallel rather than to manifest autonomous phonological representations [6]. Following this assumption, effective modeling of speech prosody can be achieved only if specific communicative functions are simulated. The present study works toward this objective by trying to derive qTA parameters that are function-specific.

Two functions are being simulated, lexical tone and focus. Tone specific parameters are derived by averaging the values of m, b and ω_n across all individual occurrences of each of the four Mandarin tones, H, R, L and F. Focus specific parameters are derived based on the findings of [10,12]. The sentences are divided into the following regions: pre-focus, on-focus, post-focus and final-focus. Final-focus is treated separately based on the findings of [10,12] as well as [4], which show that a conflict with the interrogative function forces final focus to be realized with severe compromise. Finally, for a sentence with no narrow focus, its entirety is treated as pre-focus. The parameters extracted from all focus regions are measured relatively to the average values of the pre-focus regions. For each region the averaged adjustment parameters are calculated as follows,

$$\begin{aligned} \Delta m &= \frac{1}{N}\sum_{i=1}^{N}(m_i - \bar{m}_{pre}) \\ \Delta b &= \frac{1}{N}\sum_{i=1}^{N}(b_i - \bar{b}_{pre}) \\ \Delta \omega_n &= \frac{1}{N}\sum_{i=1}^{N}(\omega_{n,i} - \bar{\omega}_{n,pre}) \end{aligned} \quad (4)$$

where Δm, Δb, and $\Delta \omega_n$ are the averaged adjustment parameters for target slope, target height, and natural frequency, respectively. N denotes the total number of syllables within that region. $\bar{m}_{pre}$, $\bar{b}_{pre}$, and $\bar{\omega}_{n,pre}$ are the means of each parameter from the pre-focus region.

5. Testing and Results

We tested the effectiveness of the new approach on the dataset obtained in [12]. We did the testing by iteratively adding more functional specifications while reducing the amount of direct resynthesis.

5.1. Dataset and method

The dataset consists of 3840 Mandarin five-syllable utterances by 4 male and 4 female speakers. In each utterance, the first and last two syllables form disyllabic words while the third syllable is a monosyllabic word. Each position has different

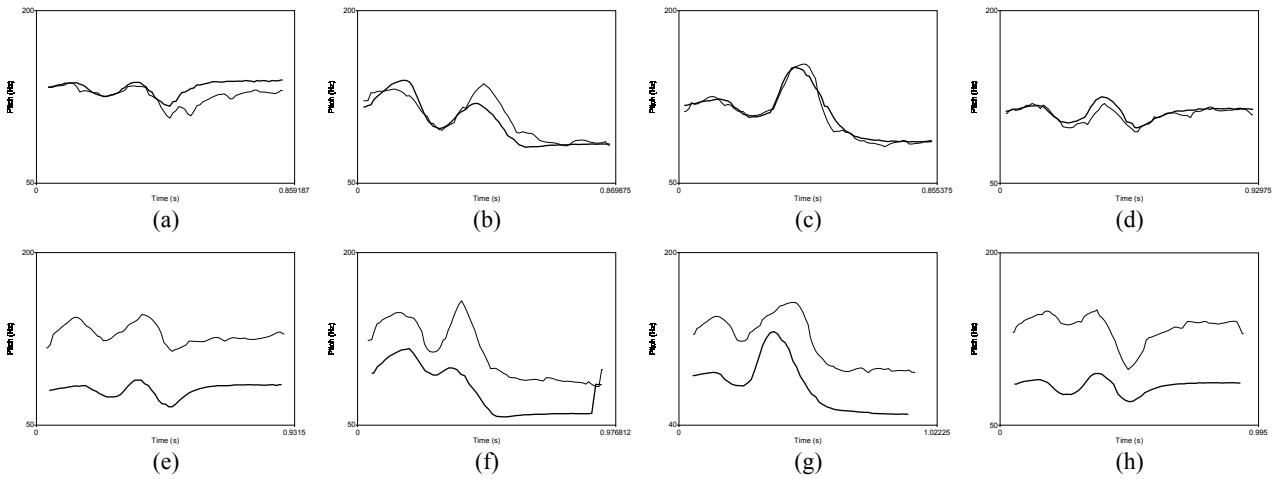

Figure 3: *The comparisons between the original F_0 (thin line) and synthesized F_0 (thick line) of the tone sequence HRFHH with* (a,e) *no focus,* (b,f) *initial focus,* (c,g) *medial focus, and* (d,h) *final focus.* (a,b,c,d) *are the results from the lowest error sentences while* (e,f,g,h) *are the results from the highest error sentences.*

Table 1: *The pre-focus parameters.*

Tone	***Male***			***Female***		
	m	***b***	ω_n	***m***	***b***	ω_n
H	0	6	21	0	-9	21
R	250	-12	18	391	-41	17
L	0	-38	20	0	-117	14
F	-474	-11	21	-710	-45	20

Table 2: *The adjustment parameters of on-focus, post-focus, and final-focus regions.*

Region	***Tone***	***Male***			***Female***		
		Δm	Δb	$\Delta\omega_n$	Δm	Δb	$\Delta\omega_n$
on-focus	**H**	0	21	0.18	0	36	-0.76
	R	229	6	0.33	354	4	-1.88
	L	0	-12	0.96	0	-44	-1.18
	F	-226	10	-1.69	-539	-3	-3.17
post-focus	**H**	0	-38	-2.89	0	-72	-3.68
	R	-25	-27	-0.09	-39	-61	2.62
	L	0	-18	-3.07	0	-48	-2.63
	F	37	-18	-1.57	31	-41	-0.78
final-focus	**H**	0	-4	-3.38	0	4	-5.27
	L	0	-11	-3.42	0	-29	-2.68

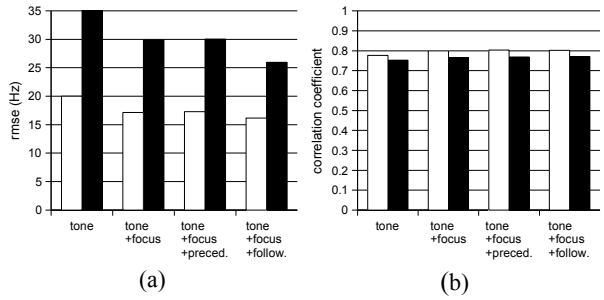

Figure 4: *The testing results, averaged (a) rmse and (b) correlation coefficient, for successively imposing more functions to the qTA model. The white and black bars represent the test results for each gender, male and female, respectively.*

tones, and each sentence has four different focus conditions: no focus, initial focus, medial focus, and final focus. For more details on the dataset, please refer to [12].

Three tests were conducted. The first was to examine the capability of the model to simulate tone-specific F_0 contours. The second test was to examine how well the model can simulate focus-specific F_0 patterns. And the third test aimed to examine the context dependency of the tone and focus functions. This was done by additionally specifying either the preceding or the following tonal context.

The tests were conducted separately for male and female speakers due to the difference in pitch range. The test candidate was circularly selected from the dataset, while the data of the rest of the speakers formed the training set. Using this selection scheme, the test repeated four times for each gender, thus maximizing the chances of detecting the worst error. The training phase began with automatically extracting the parameters from each utterance. The parameter were then averaged according to the functions to be tested and treated as the function-specific parameters in the testing phase. During the testing phase the root mean square error (*rmse*) and correlation coefficient (*r*) for each sentence were measured.

5.2.Results

Table 1 and 2 show the averaged tone-specific parameters and the corresponding focus-specific adjustment parameters, respectively. Table 1 shows the parameters learned from the pre-focus region, which also include all the words in sentences with no narrow focus, as explained earlier. The *b* values in Table 1 are measured relative to the initial F_0 value of the sentence. The parameters in Table 1 show clearly distinct target values for the tones for both male and female speakers. They agree well with the tonal targets hypothesized in [12].

Table 2 shows the averaged adjustment parameters for different focus regions expressed as differences from the parameter values in the pre-focus region shown in Table 1. These differences are consistent with the findings of [10,12]. Both *m* and *b* in the on-focus region were enhanced such that *b* is higher for H tone and lower for L tone while *m* is higher for R tone and lower for F tone. In contrast to the on-focus

region, the parameter adjustment values in the post-focus region are compressed. In the final-focus region, the *b* adjustments (Δb) are very small for H tone but slightly larger for L tone. Interestingly, the ω_n adjustments in the final-focus region are negative for both males and females. This seems to reflect the minimum target adjustments with clearly increased syllable duration in the final position [4].

Fig. 3 shows examples of comparisons between the original and synthesized F_0 with different focus conditions using the tone-specific parameters and their focus-specific adjustments from Table 1 and 2, respectively. The samples were chosen from cases with the best (top) and worst error rates (bottom). As we can see from the top plots, the model is able to predict both tones and focus well. The bottom plots show that even in the worse cases the shape of the F_0 trajectories are well simulated, while the errors mainly come from differences in overall pitch level, which are due to errors in measuring the initial values of the test sentences.

Fig. 4 shows the error and correlation results which consist of averaged *rmse* and *r* across the dataset. As expected, as more functions were specified, the error rates successively decreased and the correlation increased for both male/female speakers. Overall, *rmse* and *r* improved from 20.02/35.00 and 0.777/0.753 with only tonal specifications to 16.18/25.96 and 0.802/0.770 with focus-specific and following-tone-specific adjustments.

The results from the context dependency test shows another important characteristic of the qTA model. From Fig. 4, when the preceding context was included, there are no significant improvement in accuracy. This indicates that the qTA model already has the intrinsic ability to capture the carry-over effect. Meanwhile, including the following context considerably reduces the error rate. This is because the current version of the model does not yet include a mechanism for simulating the anticipatory dissimilation effect [12], which makes *b* of H tone higher when it is followed by L tone than by other tones.

6. Discussion and conclusion

The present study has made considerable progress toward the three objectives that we believe are critical for robust speech modeling: (a) simulating articulatory processes that are biophysically plausible, (b) automatic extraction of model parameters from natural speech, and (c) generating acoustic forms that convey specific communicative functions. For the first objective, we developed qTA [7], a quantitative version of the TA model [11], that implements a feedback-controlled second-order system, which generates F_0 contours through syllable-synchronized sequential target approximation. Being based on a set of assumptions about the biophysical mechanisms and the neuromuscular control of goal-oriented motor movements, the model is highly constrained. Only three parameters, all functionally meaningful, need to be estimated from natural speech. This has made it possible for us to achieve automatic extraction of model parameters that are applicable to not only the original but also other sentences. Finally, to generate F_0 contours that are proper for specific communicative functions, we extracted tone-specific and focus-specific parameters from a natural speech dataset based on patterns found in [4,10,12]. The testing results showed that even when applied to speakers not included in the training, the error rates were still comparable with those of previous studies [1-3,6,8]. These results demonstrate the effectiveness of the qTA model in simulating function-specific F_0 contours in speech.

Note that there are many different levels of biophysical processes. What we have tried to do through the development of the qTA model is to identify a level where the link between articulatory control and communicative functions is the most direct. The integrated force that drives target approximation can be certainly decomposed into individual muscle forces, as have been the goal of [1]. The advantage of the present approach can be seen in the finding that the extracted parameters were applicable to both original and novel sentences and even across speakers. This suggests that prosodic modeling can be done through simulation of the encoding process at the articulatory and functional level.

At the functional level, the function-specific parameter adjustments we have derived for focus can be considered as implementations of the encoding schemes in the PENTA model. We have yet to work out a quantitative system for representing different communicative functions in a unified way, however. Such a system can be developed only after many more communicative functions have been tested. Also, the robustness of our simulation needs to be further verified by perceptual tests, which we are currently planning to do.

7. References

[1] Fujisaki, H.; Wang, C.; Ohno, S.; Gu, W., 2005. Analysis and synthesis of fundamental frequency contours of standard Chinese using the command-response model. *Speech Comm. 47*, 59-70.

[2] Hirst, D.; Espesser, R., 1993. Automatic modelling of fundamental frequency using a quadratic spline function, *Travaux de l'Institut de Phonétique d'Aix 15*, 75–85.

[3] Kochanski, G.; Shih, C., 2003. Prosody modeling with soft templates. *Speech Comm. 39*, 311-352.

[4] Liu, F.; Xu, Y. *in press*. Parallel encoding of focus and interrogative meaning in Mandarin intonation. *to appear in Phonetica*.

[5] Miall, R. C.; Wolpert, D. M.; 1996. Forward Models for Physiological Motor Control. *Neural Network 9(8)*, 1265-1279

[6] Pierrehumbert, J., 1981. Synthesizing intonation. *J. Acoust. Soc. Am. 70*, 985–995.

[7] Prom-on, S.; Xu, Y.; Thipakorn, B., Quantitative target approximation model: simulating underlying mechanisms of tones and intonations. *submitted to ICASSP2006*.

[8] Taylor, P., 2000. Analysis and synthesis of intonation using the Tilt model, *J. Acoust. Soc. Am. 107*, 1697-1714.

[9] Xu, Y., 2005. Speech melody as articulatorily implemented communicative functions. *Speech Comm. 46*, 220-251.

[10] Xu, Y.; Xu, C. X., 2005. Phonetic realization of focus in English declarative intonation. *J. Phonetics 33*, 159-197.

[11] Xu, Y.; Wang, Q. E., 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Comm 33*, 319-337.

[12] Xu, Y., 1999. Effects of tone and focus on the formation and alignment of F0 contours. *J. Phonetics 27*, 55-105.