Quantitative Target Approximation Model: Simulating Underlying Mechanisms of Tones and Intonations

Santitham Prom-on\textsuperscript{1,2}, Yi Xu\textsuperscript{2} and Bundit Thipakorn\textsuperscript{1}

\textsuperscript{1}Department of Computer Engineering, King Mongkut’s University of Technology Thonburi, Thailand  
\textsuperscript{2}Department of Phonetics and Linguistics, University College London, UK

\textbf{ABSTRACT}

This paper proposes a quantitative target approximation (qTA) model for simulating tone and intonation. Based on two theoretical models: the target approximation model [13] and the PENTA model [11], the qTA model additionally incorporates several assumptions related to the underlying articulatory mechanisms, including (1) $F_0$ production can be represented by a second-order overdamped system, and (2) the system is controlled by a time-delayed feedback loop to sequentially approximate underlying pitch targets. We tested the model with the dataset from [14]. Two experiments were conducted to validate the model and to study the effect of tone, position, and focus. The results were satisfactory in term of the error rate and correlation.

\textbf{1. INTRODUCTION}

There have been many attempts to build a robust model capable of simulating tone and intonation in speech. Among them are the IPO model [4], the tilt model [9], the quadratic spline model [5], the Pirenhumbert model [7], the PalMT model [1], the command-response model [3], the soft-template model [6], and the dynamic system model [9]. Many of them try to describe the surface $F_0$ directly [1,4,5,7,9]. But a few have attempted to simulate the underlying mechanisms of $F_0$ production [3,6,8]. Among these, however, only the command-response model has tried to describe the actual physiological mechanisms of $F_0$ production [3]. In this paper we propose a new model based on a set of specific assumptions about the articulatory mechanisms of $F_0$ production. The model is a quantitative implementation of the target approximation model and is therefore referred to as the qTA model.

\textbf{2. ASSUMPTIONS}

\textbf{2.1. Second-Order Overdamping}

Following [3] we assume that $F_0$ is directly related to vocal fold tension. According to [11] and [15], there are two major muscle forces controlling vocal fold tension, namely, the tension raising force and the tension lowering force. Both forces are jointly generated by muscles that lengthen the vocal folds, the cricothyroids (CT), and those that shorten them, mainly the thyroarytenoids (TA). These muscles are antagonistic to each other, but they are jointly responsible for both raising and lowering $F_0$, although the contribution of one or the other may temporarily dominate. $F_0$ raising is due not only to the contraction of CT, but also to the simultaneous contraction of TA. Likewise, $F_0$ lowering is due not only to the contraction of TA, which shortens the vocal folds, but also to the simultaneous reduction in CT contraction. Since these muscle forces transfer energy back and forth, we may assume $F_0$ production as a second-order system. Furthermore, based on the fact that both muscle forces are controlled by the brain, as opposed to systems in which the back-and-forth energy transfer is largely compensatory, and on the fact that no observed oscillation exists in actual $F_0$, we assume that the system is overdamped, with pitch targets as the forcing function.

\textbf{2.2. Time-Delayed Feedback Control}

By definition, any goal-oriented movement by a biological organism has to be controlled with a feedback mechanism, and speech should be no exception. It would thus be reasonable to assume that feedback is a core component of a model that simulates the mechanisms of $F_0$ production in speech. There are various feedback channels potentially relevant for speech production, including auditory feedback, tactile feedback, proprioceptive feedback [15] and corollary discharge [10], among which the last is likely the fastest and the first two the slowest [12]. As our model implementation later suggests, the innermost feedback loop has to be a very fast one, as its time delay may impose a limit on the speed of articulatory movement. So we are tentatively assuming that the feedback in the qTA model is at the level of corollary discharge.

\textbf{2.3. Syllable-Synchronization}

Based on evidence discussed in [13], we assume that tones are implemented in synchrony with the syllable. In other
words, the implementation of a tone starts at the onset of the syllable and ends at the offset of the syllable, as shown in Figure 1. This assumption also implies that pitch target and other melodic primitives, such as articulatory strength and duration, are assigned locally for each syllable [11].

2.4. Sequential Target Approximation

During each target approximation, the state of articulation depends not only on the discrepancy between the current state and the target, but also on the influence of the preceding syllable [11]. As illustrated in Figure 1, at the beginning of the second syllable, while the implementation of the second target has already started, F0 is still rising due to the initial velocity and acceleration left by the first syllable. This influence, also known as carry-over effect, would gradually decrease over time. This assumption thus stipulates that the state of articulation, as defined by pitch level, velocity, and acceleration, is sequentially transferred from one syllable to the next at the syllable boundary.

3. MODEL SPECIFICATIONS

3.1. Pitch Target

According to [13], the pitch target is the minimally essential representation associated with the functional pitch events such as tones or accents. There are two kinds of pitch targets: static and dynamic. As illustrated by the dashed lines in Figure 1, a pitch target can be described with a simple linear equation,

\[ x(t) = mt + b \]  

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 scope of time, \( t \), would be relative to the onset of the syllable.

The type of pitch target reflects the characteristics of the tone. In Mandarin, H, L, and N, which have only register specifications, can be represented by static targets, while R and F, which require slope specifications, can be represented by dynamic targets. Different types of pitch targets also have different parameter ranges. The static targets have zero slope and heights that depend on the tone categories, while the dynamic targets have slopes that vary with the tone categories.

3.2. Model of the Tension Control

As discussed in section 2.1, the qTA model is based on the assumption that the tension control of the vocal folds is a second-order overdamped system. Generally, the transfer function of a second-order linear time-invariant system would have two control parameters, \( \zeta \) and \( \omega_n \), which denote the damping ratio and the undamped natural frequency, respectively. Each parameter has its own meaning with regard to articulation. \( \zeta \) indicates the responsiveness of the tension control. \( \omega_n \) indicates how much effort is used to implement the target, with greater \( \omega_n \) yielding faster approximation of the target, other things being equal.

3.3. Time-Delayed Feedback Control

The feedback control mechanism used in the model is illustrated in Figure 2. To account for the effect of time delay on the system and to find a closed form of the solution at the same time, we approximate the time delay using the first-order Padé approximant which increases the order of the overall model by one. Thus, the complete response of the model is in the third-order form.

\[ \begin{align*}
\text{Pitch target} & \quad x(t) = mt + b \\
\text{2-nd order model} & \quad \\
\text{Tension control} & \quad \zeta, \omega_n \\
\text{Time delay} & \quad T_d
\end{align*} \]

Figure 2. A block diagram of the feedback control scheme used in the qTA model. The double preamplifier is used to compensate for the halving reduction needed for system stability and to guarantee that the forcing function resembles the pitch target.

3.4. Syllable-Synchronized Sequential Implementation

The syllable-synchronized sequential target approximation was extensively discussed in [11] and also in section 2.3 and 2.4. The target and model parameters are assigned locally in each syllable. At the inter-syllable boundary, the articulatory state is transferred from the preceding syllable to the current syllable as the initial conditions.

\[ \begin{align*}
F_0(0) &= F_0(t_{i-1}^{\text{final}})_{i-1} \\
F_{\phi}(0) &= F_{\phi}(t_{i-1}^{\text{final}})_{i-1} \\
F_{\varphi}(0) &= F_{\varphi}(t_{i-1}^{\text{final}})_{i-1}
\end{align*} \]  

(2)

This transferring mechanism reflects the propagation of the laryngeal state across the syllable boundary [11].

Figure 1. An illustration of the theoretical pitch target approximation model proposed in [13]. In each syllable (demarcated by the vertical lines), \( F_0 \) approaches the pitch target asymptotically. The dashed lines represent the pitch targets while the solid curve represents the surface \( F_0 \).
3.5. $F_0$ Realization

We can derive the ordinary differential equation by combining the transfer function of the second-order model with time-delayed feedback. The solution of the differential equation consists of two parts: the forced response and the natural response. Our model assumes that at the end of a complete execution in each syllable, i.e. when the time of the execution approaches infinity, $F_0$ would reach the target. Therefore, the complete realization of the model would be

$$F_0(t) = c_1e^{r_1t} + c_2e^{r_2t} + c_3e^{r_3t} + mt + b \quad (3)$$

where $r_1$, $r_2$, and $r_3$, are the roots of the homogeneous equation of the differential equation. The coefficients $c_1$, $c_2$, and $c_3$ are solved from the initial conditions.

4. EXPERIMENTS

To test the effectiveness of the model, we conducted a series of experiments using a dataset collected in [14]. The goal was to obtain the model parameters through training and then test how well the model-generated $F_0$ contours fit the actual contours.

4.1. Dataset

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 are disyllabic words while the third syllable is a monosyllabic word. The first and the 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 forth syllable. To avoid the tone sandhi which changes LL to RL, the L tone was excluded from the tone set of the third 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. Further details on the dataset can be found in [14].

We divided the dataset into two parts for training and testing purposes. The training set consists of speech by 3 male and 3 female speakers (2880 utterances) while the testing dataset consists of speech by 1 male and 1 female speakers (960 utterances).

4.2. Estimating Model Parameters

To estimate the parameters of the qTA model, we varied each parameter within the search space and evaluated the sum square error in each syllable. We then obtained the parameter with lowest error. The sum square error was defined as

$$E = \sum_{n} (f_{0}(n)_{\text{act}} - f_{0}(n)_{\text{sys}})^2 \quad (4)$$

For target slope, the search space depended on the type of the pitch target, i.e., whether it was static or dynamic. The search space for target height was centered around the final $F_0$ value with a small vertical range. The damping ratio and the time delay were fixed based on initial calibration. The parameter estimation was done for each sentence separately. Thus focus was not given any special treatment at this stage. Table 1 shows the averaged trained model parameters separated by gender.

<table>
<thead>
<tr>
<th>Tone</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$m$</td>
<td>$b$</td>
</tr>
<tr>
<td>H</td>
<td>0.00</td>
<td>0.49</td>
</tr>
<tr>
<td>R</td>
<td>399.39</td>
<td>-18.05</td>
</tr>
<tr>
<td>L</td>
<td>0.00</td>
<td>-47.86</td>
</tr>
<tr>
<td>F</td>
<td>-579.38</td>
<td>-2.06</td>
</tr>
</tbody>
</table>

Table 1. Averaged model parameters using $\zeta=1.5$ and $T_d=0.005$ s. The value of $b$ is relative to the initial $F_0$ of the sentence.
4.3. Experiment 1: Tone and Position Specific Testing

We first tested the model by using the averaged parameters as shown in Table 1 to predict the $F_0$ of the test set and computed the root mean square error (rmse in, Hz) and the correlation coefficient ($r$). The results are shown in the left half of Table 2. While these results are not bad, further improvement was seen when we averaged the parameters of each tone separately for each of the five positions in the sentence, as shown in the right half of Table 2.

Table 2. Comparative results between tone specific only testing and tone and position specific testing. Focus is ignored in this experiment.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Tone specific only</th>
<th>Tone and position specific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rmse</td>
<td>r</td>
</tr>
<tr>
<td>Male</td>
<td>17.76</td>
<td>0.791</td>
</tr>
<tr>
<td>Female</td>
<td>32.91</td>
<td>0.772</td>
</tr>
</tbody>
</table>

4.4. Experiment 2: Focus Specific Testing

As found in [14], focus introduces pitch range adjustments around the focused item, expansion at focus, suppression after focus and little change before focus. Thus simulating focus with the qTA model could be done by separately adjusting the model parameters in different regions [11]. We tested this possibility by separately averaging the pre-focus, focus, post-focus, and final focus parameters. Improvement can be seen in Table 3 as compared to Table 2. Initial results from an on-going perception experiment also confirmed the effectiveness of the synthesized focus.

Table 3. Results of including the focus and position in testing.

<table>
<thead>
<tr>
<th>Gender</th>
<th>rmse</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>13.74</td>
<td>0.819</td>
</tr>
<tr>
<td>Female</td>
<td>28.45</td>
<td>0.787</td>
</tr>
</tbody>
</table>

5. DISCUSSION AND CONCLUSION

The qTA model is proposed based on a set of specific assumptions from our understanding of both the biophysical constraints of $F_0$ articulation and the neural control of goal-oriented motor movements. Like [3], we assume that $F_0$ changes originate from muscle contractions. But we further assume that muscles act in functional groups [15] rather than each as separate sub-systems [3]. We also assume that the articulatory mechanism is an overdamped system driven by a forcing function because the active control of both the agonist and antagonist muscles makes it more responsive than what is assumed in [3]. We further assume that $F_0$ production is a neurally controlled process of syllable-synchronized sequential target approximation, and as such it can be simulated as a feedback controlled second order system that successively approximates locally defined pitch targets, each within the time scope of a syllable.

The assumptions of the qTA model make it more restrictive than most other models, but they also make it economical. Only three parameters need to be optimized: slope ($m$) and height ($h$) of the targets and natural frequency ($\omega_n$) of the second order system; and they are all meaningful in terms of communicative functions in speech [11]. This makes the model an effective tool not only for testing theories of tone and intonation, but also for generating $F_0$ contours in speech synthesis, as our preliminary synthesis results have demonstrated. Although we cannot directly compare the results with other works due to the different datasets used, the qTA model performed quite compatible to both direct $F_0$ specification models [1,2,9] and other articulatory-oriented models [3,6,8]. The concepts behind the model are also potentially useful for improving speech recognition systems and for testing theories of speech perception.

7. REFERENCES