A Simplified Method of Learning Underlying Articulatory Pitch Target

Hao Liu, Yi Xu

Department of Speech, Hearing and Phonetic Sciences, University College London, UK
{h.liu.12, yi.xu}@ucl.ac.uk

Abstract

Previous research has shown that parameters of the quantitative Target Approximation model (qTA) proposed by Prom-on and Xu can be directly extracted from natural speech with high accuracy through analysis-by-synthesis implemented in PENTA-trainers. While this may raise the possibility that PENTAtrainers actually simulate natural acquisition of prosody production, it is questionable that the human brain actually replicates the full articulatory mechanics represented by qTA in order to learn and control prosody production. In this paper we explore if a much simpler function can be used to extract at least some of the qTA parameters. We first managed to reduce the number of qTA parameters from three to two by evaluating their relative sensitivity. We then tested a pursuit function that learns only pitch target height and slope. Using a corpus of Mandarin utterances varying in lexical tone and focus, we show that parameters learned by the pursuit function can be used in qTA synthesis to generate F0 contours closely resembling those generated with parameters learned with qTA-based analysis-by-synthesis, with the advantage of having a much simpler learning algorithm. These results suggest that it is possible to learn articulatory control parameters for prosody without fully replicating the mechanical process itself.

Index Terms: F0 contour modelling, target approximation, pursuit curve

1. Introduction

It has been recently demonstrated that F0 contours closely resembling those of natural speech can be generated by the PENTA model [1] with a small number of functionally specific pitch targets extracted directly from raw speech data [2, 3]. The F0 contour generation in those studies is done by the quantitative target approximation model (qTA), which simulates a third-order linear system [2]. Two automatic algorithms have been developed, as implemented in PENTAtrainer1 [4] and PENTA-trainer2 [5], to extract the parameters of qTA model from functionally annotated speech data using analysis-by-synthesis controlled by either exhaustive [4] or stochastic [5] optimizations. Such parameter extraction processes could be imagined as analogous to the natural speech acquisition process in which the child presumably learns to speak by discovering, also through analysis-by-synthesis [6], the articulatory control parameters needed to generate adult-like speech patterns. There are two potential problems with this analogy, however. The first is that the number of analysis-by-synthesis cycles is unrealistically large. The second problem is that the analogy assumes that either the child overtly imitates the same adult utterance over and over again, or develops a virtual replica of the qTA model in the brain for both learning and controlling the production of tone and intonation. With these problems in mind, in this paper we explore an alternative learning mechanism that a) uses a simplified model that approximates the core properties of qTA, and b) does not require analysis-by-synthesis searching process.

We will first try to reduce the model complexity from qTA by reducing the number of parameters from three to two by comparing the relative sensitivities of the three model parameters of qTA. We will then test a “pursuit” function [7], using a corpus of Mandarin utterances varying in lexical tone and focus, to show that the pursuit function can learn the two remaining target parameters directly, with the learned values very similar to those found by exhaustive analysis-by-synthesis as implemented in PENTAtrainer1.

2. Pitch modeling

2.1. Target Approximation model

The quantitative target approximation model (qTA) assumes that continuous surface F0 contours are the results of successive, yet non-overlapping underlying articulatory movements, each approaching an underlying target associated with a local host syllable. A target can be either static or dynamic (Figure 1), which can be represented by a simple linear equation:

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

where b is target height, m is target slope and t is time relative to the onset of the host syllable.

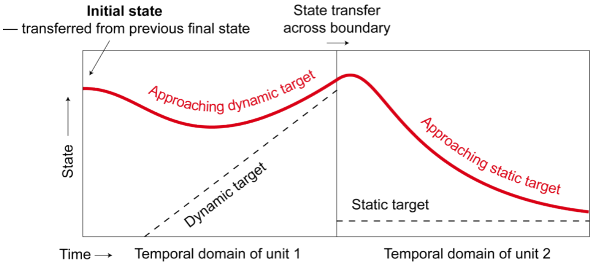
Initial state
— transferred from previous final state
State transfer
across boundary
State
Approaching dynamic target
Dynamic target
Approaching static target
Static target
Time →
Temporal domain of unit 1
Temporal domain of unit 2

Figure 1: *Target approximation model.*

The qTA model is a third-order critically damped linear system as represented by the following equation

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

where $f_0(t)$ is the complete form of the fundamental frequency in semitones, $x(t)$ is the forced response and the polynomial and the exponential are the natural response [2]. λ is the rate of target approximation, i.e., how rapidly the target is approached. The transient coefficients c_1, c_2 and c_3 are jointly determined by the initial F0 dynamic state of the syllable, consisting of F0

10.21437/SpeechProsody.2014-192

level, velocity, and acceleration transferred from the offset of the preceding syllable:

$$c_1 = f_0(0) - b, \tag{3}$$

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

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

At the end of the syllable, the final F0 dynamic state is transferred to the next syllable to become its initial state, which results in a smooth and continuous F0 trajectory across the syllable boundary.

2.2. Control parameter sensitivity assessment

In order to find a simple model that can approximate the qTA model we examined the sensitivity of F0 contours generated by qTA to variations in the three pitch target parameters: m, b, λ.

A six syllable Mandarin phrase, /wó yǒu yí wèi yá yī/, was chosen and recorded by a male native speaker of Mandarin. PENTAtrainer1 [4] was used to find an optimal combination of m, b and λ. Then, three sets of F0 contours were generated by varying one parameter while holding the other two constant at their optimal values. The difference between the generated contour and the optimum contour was then analysed in terms of semitone shifts, as described in the paragraphs below.

Figure 2 displays the error vectors for which the target slope m and TA rate λ are set to be the same as the optimal values but target height b is given five different values. The graph is from the rising tone syllable /wó/. The pattern distributions of error vectors are very regular — all the curves gradually move away from x-axis at the same pace. The common starting point is because qTA is a sequential model, and all variations in the current syllable step from the same offset value of the previous syllable. Because /wó/ is the first syllable of the chosen utterance, the starting point is always zero. When the values of b are equidistant from each other, the error vector curves exhibit an even distribution. Note that the middle curve represents the error vector resulting from subtracting the natural F0 contour of the syllable from the contour generated with the "optimal" parameters. The very small deviations from the x-axis indicates that the two contours are very similar to each other.

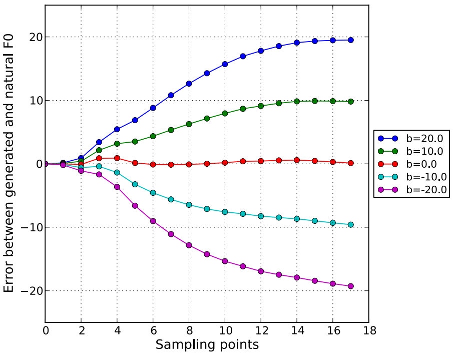

Figure 2: *Error vectors with varying b while m and λ are the same as the optimal values. Measured in semitone.*

Figure 3 displays error vectors of contours that vary in m while b and λ are held constant at their optimal values. The fusiform shaped distribution here is due to the fact that in qTA b is defined as the ending point of a target. As a result, all the generated contours have a fixed tail height by the end of the syllable, i.e., they shared the same offset. This means that, when b is held at its optimal value, the offset F0 of a syllable is virtually guaranteed to be near optimum.

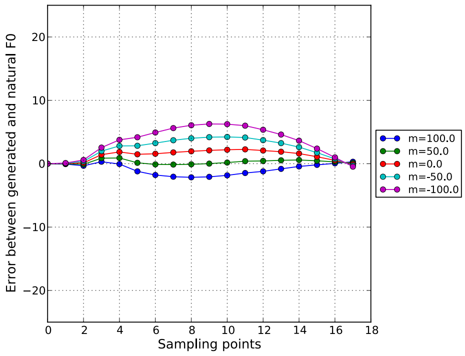

Figure 3: *Error vectors with varying m while b and λ are the same as the optimal. Measured in semitone.*

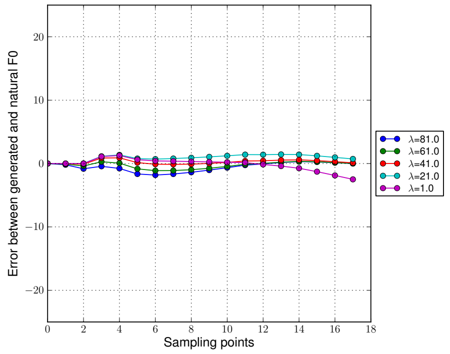

Figure 4: *Error vectors with varying λ while m and b are the same as the optimal. Measured in semitone.*

Figure 4 displays error vectors of cases where only λ is set to vary from the optimal value. The deviations are very small, indicating a much weaker effect than those of b and m. However, this does not mean that λ is unimportant in all cases. It has in fact been demonstrated that, when modelling data contain unstressed syllables and the neutral tone, the role of λ is crucial [3, 8].

The conclusion is that among the three qTA parameters, λ, i.e., target approximation rate, is less important than b, target height and m, target slope, at least for the present data set. It further suggests that a learning procedure that can find close-fitted target heights and target slopes may provide a good approximation to qTA, especially for data where λ does not have important functional significance.

2.3. Pursuit functions for pitch target estimation

The qTA model was designed to generate contours from underlying pitch targets. It is a third-order differential equation which contains non-linear elements. In this form it is not easy for learning, i.e., finding optimum underlying pitch targets from input signal. There is no analytical expression for the inverse of qTA, and so analysis-by-synthesis has to be used to estimate the model parameters from natural speech data. To make mathematical inversion possible, we investigated a method involving “pursuit” functions.

A pursuit curve is the path of an object that seeks to pursue another moving object. Consider a simple case of a hound chasing a fox, where the fox is moving at constant speed and constant direction. The pursuit curve is found under the assumption that the direction of the hound is always towards the current location of the fox, i.e. that the tangent of the pursuit curve at time t is directed towards the location of the hound at time t. Figure 5 shows an example pursuit curve.

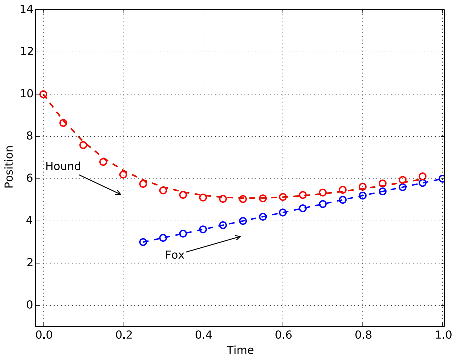

Figure 5: *A pursuit curve. Arrows indicate velocity directions.*

In the situation shown in Figure 5, where the pursued object is moving at constant velocity, the pursuit curve can be shown to have the form of an analytic equation

$$H(t) = F_0 + vt + (H_0 - F_0)e^{-t/l}, \tag{6}$$

where H_0, F_0, v, t, l denotes the initial position of the hound, initial position of the fox, velocity of the fox, time series and the “time lead” of the fox, respectively. We can interpret this as the pursuer attaining the location and velocity of the pursued according to some exponentially decreasing value of time. The rate of attainment is simply related to the time lead of the pursued.

If we use a pursuit curve to simulate the target approximation process, the linear path of the pursued becomes the underlying pitch target, the pursuit curve is the observed F0 contour, and the initial velocity of the pursuer becomes the initial conditions for the F0 at syllable onset.

As a simpler target estimation function, the pursuit function itself does not fit F0 contours as cleanly as qTA since it allows for instantaneous changes in velocity and acceleration at syllable boundaries. This can lead to rather unnatural looking F0 contours as the pursuer changes from one pursued target to another, as can be seen in Figure 6.

Since, however, we only want a simple way of deriving pitch targets represented by b and m from the input contours, these discontinuities may not be important. This is because, once the targets are learned, we will still use qTA to generate the contours from the targets, free of discontinuities at syllable boundaries.

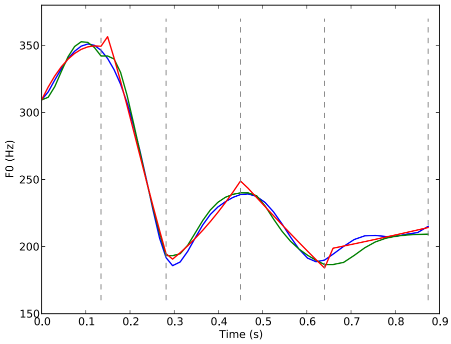

Figure 6: *Clear sharp turning points of pursuit curve at syllable boundaries. (Natural contour in blue, qTA fitted contour in green, pursuit curve in red.)*

In the next section we will compare the pursuit function with qTA-based analysis-by-synthesis in terms of quality of fit using a small corpus of utterances.

3. Pitch target learning

3.1. Data

480 utterances recorded from a female native Mandarin Chinese speaker were used to explore the fit of the target approximation model. This corpus was originally collected to examine the effects of lexical tones and focus on the formation and alignment of F0 contours [9]. The corpus consists of 24 sentences, each of them was said with four different focus locations and repeated five times. Every sentence consists of three Chinese words, the first and the third are bisyllabic and the second is monosyllabic. So there are five syllables in each sentence (Table 1). Further, the second, third and fourth syllables have varying lexical tones, which were the target syllables for the current experiment. The four focus conditions are: neutral focus (no focus), initial focus (on word 1), medial focus (on word 2) and final focus (on word 3). When a syllable is on-focus, its preceding syllable is pre-focus and its following syllable is post-focus.

Table 1: Tone patterns and corresponding sentences used as recording material. H, R, L, and F represent high, rising, low, and falling tones, respectively.

Word 1	Word 2	Word 3
HH māomī	H mō	HH māomī
HR māomí	R ná	LH mǎdāo
HL māomǐ	F mài	
HF māomì		

3.2. Method

The goal of the experiment was to learn the optimal pitch target slope and height, for each lexical tone in each focus condition, using pitch targets learned by both qTA-based analysis-by-synthesis and the pursuit curve function. Results are then compared in terms of similarity of the discovered targets, overall quality of fit.

Both methods were implemented in Python. In the case of qTA, the exhaustive local search algorithm proposed by [2] was used, with rate of target approximation (λ) held constant at 41.0. The algorithm read the data and parameter constraints, and then iteratively tested all combinations of target values using a set of possible values for target height and target slope for each utterance separately. The parameters that showed the lowest sum square error between the generated and natural F0 contours for each utterance were chosen as the target. The optimal targets for each utterance were then averaged to derive different targets for each tone and focus condition.

For testing the pursuit function, the time lead of the pursued (l) was fixed at 0.075s. Like the qTA approximation rate parameter, the pursued time lead controls the rate at which approximation takes place and might be considered a characteristic of the speaker or speaking style [8, 10].

To fit the pursuit function, a linear least squares method was used over the whole data set. Each observed F0 measurement was expressed in terms of a number of coefficients applied to a vector of 32 unknowns, being the target height and slopes of the 4 tones in the 4 focus conditions. The least squares fit derives the values of the 32 unknowns that minimise the squared error of prediction of the data by the model.

3.3. Results and evaluation

The value of the pursuit function is that it provides a direct means to determine the optimal pitch targets from the measured F0. We would still like those targets to be compatible with qTA, since as mentioned above, the pursuit function has some intrinsic inadequacies for F0 contour generation.

Table 2: Learned functional pitch targets. For focus function, PRE, ON, and POS stand for pre-focus, on-focus, and post-focus regions, respectively.

Focus	Tone	Target slope (st/s)		Target height (st)	
		qTA	pursuit	qTA	pursuit
PRE	H	2.1	7.4	19.0	18.2
	R	30.7	17.7	11.6	12.9
	L	-16.4	-40.9	16.5	20.1
	F	-29.5	-27.3	21.5	21.6
ON	H	2.5	3.3	19.4	19.3
	R	28.7	12.5	10.8	13.0
	L	-79.1	-108.4	21.6	27.0
	F	-56.4	-43.2	28.3	27.1
POS	H	-11.6	-16.8	16.4	17.3
	R	11.9	-1.9	11.3	13.1
	L	-79.7	-119.0	20.4	27.3
	F	-36.8	-35.5	20.1	20.4

As shown in Table 2, the pitch target heights found using the pursuit function were very similar to those found using qTA. There are greater differences between the estimated pitch target slopes, but they are broadly of similar sign and size.

Figure 7 illustrates F0 contours generated from the qTA targets and pursuit targets shown in Table 2 for an utterance containing a “LHL” syllable sequence with focus on the second syllable. The generated contours for qTA and the pursuit function are similar to each other and to the measured F0.

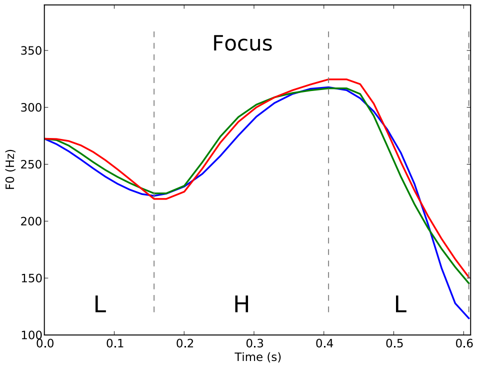

Figure 7: *Contour generated with pursuit targets is very close to the one with qTA targets. (Natural contour in blue, contour generated with qTA targets and pursuit targets are in green and red, respectively.)*

The quality of the targets obtained by the two methods was evaluated by measuring the root-mean-square error (RMSE) of prediction of the data set using the qTA model and the discovered targets. For the targets found by qTA-based exhaustive search, the RMSE of prediction is 1.83 semitones (Pearson $r = 0.8453$). For the targets found by the pursuit function, the RMSE is 1.81 semitones (Pearson $r = 0.8458$). Thus the new method of estimating underlying targets is as least as good as using the qTA model directly for inversion.

In terms of learning efficiency, although the linear least squares with pursuit function can get target parameters in a blink of time, we didn’t expect it to be a replacement of simulated annealing as implememnted in PENTAtrainer2. Instead, we hope in future work, it is possible that further improvements can be made, perhaps by seeding the qTA model with pursuit function found targets and then applying some iterative hill-climbing method to find a local minimum error of prediction.

4. Conclusion

In this study, we have shown how a pursuit function can be used in place of the qTA function for the problem of finding underlying pitch targets from measured F0. From the results we can see that the pursuit function enabled a direct means of finding the pitch targets, and more importantly, that the learned targets could be reused by the qTA model for F0 contour production. This finding provides support for our hypothesis that without fully replicating the mechanical process itself, articulatory control parameters for prosody can be learned with simpler learning process which is more conceivably developed in motor control by the human brain.

5. Acknowledgments

We would like to thank Dr. Mark Huckvale (UCL) for originally suggesting the study. The content of the paper has been greatly improved by his helpful comments.

6. References

[1] Xu, Y., “Speech melody as articulatorily implemented communicative functions,” *Speech Communication*, vol. 46, no. 3, pp. 220–251, 2005.

[2] Prom-on, S., Xu, Y., and Thipakorn, B., “Modeling tone and intonation in Mandarin and English as a process of target approximation,” *The Journal of the Acoustical Society of America*, vol. 125, p. 405, 2009.

[3] Xu, Y. and Prom-on, S., “Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning,” *Speech Communication*, vol. 57, pp. 181–208, 2014.

[4] Xu, Y. and Prom-on, S., “PENTAtrainer1.praat. Available from: http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/.”

[5] Prom-on, S. and Xu, Y., “PENTAtrainer2: A hypothesis-driven prosody modeling tool,” *ExLing 2012*, p. 93, 2012.

[6] Stevens, K. N. and Halle, M., “Remarks on analysis by synthesis and distinctive features,” *Models for the Perception of Speech and Visual Form*, pp. 88–102, 1967.

[7] Boole, G., *A treatise on differential equations*. Macmillan & Company, 1859.

[8] Prom-on, S., Liu, F., and Xu, Y., “Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling,” *The Journal of the Acoustical Society of America*, vol. 132, p. 421, 2012.

[9] Xu, Y., “Effects of tone and focus on the formation and alignment of F0 contours,” *Journal of Phonetics*, vol. 27, no. 1, pp. 55–105, 1999.

[10] Xu, Y., “Fundamental frequency peak delay in Mandarin,” *Phonetica*, vol. 58, no. 1-2, pp. 26–52, 2000.