A PITCH TARGET APPROXIMATION MODEL FOR F_0 CONTOURS IN MANDARIN

Ching X. Xu[*], Yi Xu[*], and Li-Shi Luo[†]

[*]Northwestern University, Evanston, IL, USA, [†]NASA Langley Research Center, Hampton, VA, USA

ABSTRACT

In this paper we report a quantitative model for generating F_0 contours in Mandarin Chinese. The model assumes that a) the basic units underlying Mandarin tones are pitch targets that are either static or dynamic, b) each pitch target is assigned to and implemented in synchrony with a syllable, and c) throughout the duration of a syllable, the pitch target is implemented continuously and asymptotically. The model expresses the F_0 curve in each syllable as an exponential asymptote that approximates an underlying pitch target. Parameters in the model specify a) the properties of the pitch target, including its height and direction of movement, b) the rate at which the target is approximated, and c) how smoothly two adjacent tones are joined at the syllable boundary. Preliminary effort to fit the model to existing F_0 data has yielded encouraging results.

1. INTRODUCTION

Xu and Wang [4] proposed a theoretical framework for accounting for F_0 contours in Mandarin, and potentially for F_0 contours in other languages as well. The framework is based on a set of explicit assumptions inspired by recent findings about F_0 contours in Mandarin [1, 2, 3]. The essential assumption of the framework is that observed F_0 contours are not linguistic units per se. Rather, they are the outcome of implementing linguistically functional units such as tone and pitch accent under various articulatory constraints. Similar to segmental phonemes, tone and pitch accent are also abstract units, underneath which are articulatorially operable units comparable to segmental phones. These units are referred to as pitch targets.

The second assumption is that there are two basic kinds of pitch targets — static and dynamic. A static pitch target has a register specification, such as high, low, or mid. A dynamic pitch target has a linear movement specification, such as rising or falling.

The third assumption is that each pitch target is assigned to a host unit, such as a syllable; and it is implemented in synchrony with the host unit, i.e., starting at its onset and ending at its offset.

The fourth assumption is that, throughout the duration of the host, the approximation of a pitch target is continuous and asymptotic.

The framework in its original form is qualitative, however. Quantitative implementation of the framework is needed to test both its theoretical adequacy and its applicability to speech synthesis and recognition. The present paper reports the results of our preliminary attempt to quantify the Xu and Wang framework [4].

2. THE MODEL

2.1. The Form of the Model

The production of F_0 contours is known to be controlled by neural commands issued to laryngeal muscles such as cricothyroid, sternohyoid, thyroarytenoid, etc. [5]. However, these muscles are used either as agonists or antagonists, depending on factors like tonal contexts, segmental units concurrently produced with the tone, and individual differences in laryngeal structure. It would be therefore rather difficult to generate proper F_0 contours by directly simulating the activities of the larynx. On the other hand, the F_0 data reported in recent studies [1, 2, 3] indicate that regardless of what exactly the laryngeal activities are, the end result is to generate F_0 contours that seem to asymptotically approach simple underlying pitch targets. Our model therefore is designed to simulate the effects of the aggregated force of the laryngeal controls rather than the effects of individual neural commands.

We assume that the observed asymptotic approximation of a target is due to an aggregated force in the form of a dissipative function

$$F = -\lambda v = -\lambda dy / dt \qquad (1)$$

where λ is a positive constant and v is the velocity of approximating a target.

According to Newton's second law of motion,

$$F = ma \qquad (2)$$

where F, m and a are the force acting on a body, the mass of the body, and the acceleration of the body, respectively, we have

$$-\lambda dy / dt = ma = md^2y / dt^2 \qquad (3)$$

Solving (3) gives us an exponential function:

$$y = \alpha \bullet exp(-\lambda t / m) + \beta \qquad (4)$$

where y is the distance from the target, t is time, and α and β are constants. Letting β be 0, we then have

$$y = \alpha \bullet exp(-\lambda t / m) \qquad (5)$$

From (5) we can get

$$y(t) = \alpha \bullet exp(-\lambda t / m), \qquad (6)$$

and

$$y(t+1) = \alpha \bullet exp(-\lambda(t+1) / m). \qquad (7)$$

If we let

$$k = exp(-\lambda / m), \qquad (8)$$

then (7) becomes

$$y(t+1) = ky(t). \qquad (9)$$

Since λ and m are both positive, k is a positive constant between 0 and 1. The value of k determines the rate of decay: the smaller its value, the faster the decay. The rate of decay can be considered as corresponding to the amount of effort the speaker uses to reach the target: the greater the effort, the faster the decay.

(9) expresses the relationship between the value of y at a given time instant t and its value at the next time instant $t + 1$. Recall that y is the distance from the target. We can rewrite $y(t)$ as ($y(t)$ – C), where C is the pitch target to be reached, and $y(t)$ is F_0 in semitone. We thus arrive at the following exponential decay function for the F_0 contour in each syllable:

$$y(t+1) = k(y(t) - C) + C; \tag{10}$$

The target C can be either a constant, when the tonal target is static, such as high and low, or a linear function, when the tonal target is dynamic, such as rising and falling. A dynamic target is expressed as

$$C = at + b; \tag{11}$$

where a is the slope of the line, and b is the intercept of the function at $t = 0$. a is positive when the target is rising, and negative when it is falling. Here we can see that a static target is in fact a special case of (11) in which a is 0 and b is the target.

Due to the continuous nature of pitch production [3], the starting F_0 in a syllable with an initial sonorant is equivalent to the ending F_0 of the preceding syllable. Therefore, in a sequence of syllable 1+syllable 2,

$$y2_0 = y1_{end} \tag{12}$$

where $y1_{end}$ is the final F_0 of syllable 1 and $y2_0$ is the initial F_0 of syllable 2 .

We further assume that the transition between the F_0 contours of two successive syllables is smooth (given that there is continuous voicing at the boundary) due to the constraints of the larynx. Hence, an additional function is needed to join the F_0 contours of adjacent syllables. A pair of hyperbolic tangent functions is thus used, and the resulting function for the connected F_0 curve of two successive syllables becomes:

$$\begin{aligned} f(t) = {} & 1/2 \bullet (1 - \tanh(mt_3) f_1(t_1) \\ & + 1/2 \bullet (1 + \tanh(mt_3) f_2(t_2) \end{aligned} \tag{13}$$

where f_1 and f_2 are F_0 functions for syllables 1 and 2, respectively; t_1, t_2 and t_3 are the time for f_1, f_2 and the hyperbolic connection function; and m determines the contributions of f_1 and f_2 to f extended across the syllable boundary. The greater the value of m, the shorter the duration of the extended contribution.

2.2. Fitting the Model

To find the parameter values for the model, a nonlinear regression procedure in SPSS was used. The procedure fits the model to the raw data by minimizing the mean square error between the fitted values and raw data. The raw data were the mean F_0 curves of disyllabic sequence /mama/ carrying all combinations of the four Mandarin tones (averaged across 8 males speakers of Beijing Mandarin) [1]. The F_0 values were first converted to semitones, with the minimum F_0 in the entire data set (65 Hz) as the base reference (i.e., semitone = 0 at 65 Hz). In the first step of curve fitting, none of the parameters was pre-fixed. Table 1 shows the parameters obtained through curve fitting. The goodness of fit is indicated by r^2. Table 1 demonstrates that the model can be well fitted to the mean F_0 curves of bitonal sequences in Mandarin if the parameters are allowed to vary.

	Syllable 1			Syllable 2				
	a_1	b_1	k_1	a_2	b_2	k_2	m	r^2
HH	—	11.58	.844	—	11.65	.801	.11	.968
HR	—	10.96	.738	1.00	-46.30	.973	.12	.968
HL	—	11.62	.791	—	-18.26	.983	.12	.995
HF	—	12.06	.888	-.97	48.93	.968	.47	.971
RH	.68	-21.62	.971	—	11.59	.950	.10	.997
RR	.86	-30.60	.971	.95	-42.81	.974	.11	.969
RL	1.00	-28.95	.963	—	-27.84	.988	.09	.989
RF	.70	-20.89	.970	-1.14	54.40	.964	.11	.993
LH	—	4.09	.917	—	12.18	.926	.14	.993
LR	—	1.08	.949	.68	-23.74	.976	.80	.977
LL	1.05	-30.78	.962	—	-29.60	.989	.09	.987
LF	—	4.75	.902	-1.51	84.07	.970	.13	.979
FH	-.62	33.14	.962	—	11.23	.900	.13	.995
FR	-.72	30.12	.950	.98	-50.44	.979	.80	.992
FL	-.74	30.23	.944	—	-9.01	.982	.24	.999
FF	-.57	34.28	.965	-.99	46.64	.961	.12	.994

Table 1. Parameters and r^2 for fitting the model as specified in (13) to the raw data. H, R, L and F represent the High, Rising , Low and Falling tones, respectively.

Since the ultimate goal of the model is to accurately predict F_0 contours with a small set of fixed parameters when given sufficient high-level information, we need to be able to reduce the ranges of all the parameters. We first noticed that the value of m is similar in most cases and the model remains well fitted when m varies from about .1 to 1 (when other parameters were allowed to vary). We therefore fixed m at its grand mean in Table 1 (m = .23). The results of re-fitting the model with the fixed m are presented in Table 2.

	Syllable 1			Syllable 2				
	a_1	b_1	k_1	a_2	b_2	k_2	m	r^2
HH	—	11.65	.880	—	11.65	.880	.23	.964
HR	—	11.67	.807	.98	-47.52	.975	.23	.961
HL	—	12.19	.834	—	-32.26	.989	.23	.993
HF	—	12.22	.900	-1.50	95.77	.980	.23	.977
RH	.51	-8.99	.961	—	11.39	.917	.23	.995
RR	.37	-2.01	.964	.84	-38.53	.975	.23	.936
RL	.64	-12.56	.961	—	-112.73	.997	.23	.977
RF	.92	-30.19	.969	-1.28	62.75	.967	.23	.993
LH	—	3.38	.914	—	12.34	.929	.23	.991
LR	—	1.96	.938	.74	-29.35	.978	.23	.977
LL	.85	-24.68	.967	—	-134.80	.997	.23	.974
LF	—	3.86	.897	-1.57	90.79	.972	.23	.974
FH	-.23	14.52	.931	—	11.34	.913	.23	.981
FR	-.53	22.58	.939	.82	-35.84	.976	.23	.992
FL	-.85	35.42	.948	—	-6.07	.978	.23	.999
FF	-.22	16.03	.942	-1.24	65.80	.970	.23	.974

Table 2. Parameters and r^2 for fitting the model to the raw data with fixed m.

If we assume that the pitch target and the effort to reach the target remain constant for each tone, we should be able to fixed a,

b, and k, for each tone. We thus attempted to fix the values of these parameters by using their means in each tone, as shown in Table 3. The only exception is that the value of b for L is fixed at -1 rather than at its mean of -39.52, because we assume that the real target of L should not be too much lower than the minimum F_0 in the raw data. The r^2 values for fitting the model with the parameters in Table 3 are shown in Table 4.

	a	*b*	*k*	*m*
H	—	11.81	.883	.23
R	.74	-25.52	.970	.23
L	—	-39.52(-1)	.938	.23
F	-.93	50.96	.967	.23

Table 3. Fixed parameters based on Table 2.

HH	HR	HL	HF	RH	RR	RL	RF
.875	.565	.934	.140	.838	.146	.728	.837
LH	LR	LL	LF	FH	FR	FL	FF
.642	.290	.744	.516	.167	.692	.884	.017

Table 4. r^2 values for fitting the model using the parameters in Table 3.

As can be seen in Table 4, while many of the r^2 values remain fairly high, some have become unacceptably low. Reexamination of Table 2 reveals that the parameters for each tone are quite different at different syllable positions. We therefore took the means of the parameters for syllables 1 and 2 separately. However, while the goodness of fit in most cases improved substantially, the two dynamic tones, namely, R and F, remained problematic in syllable 2 in some cases. We thus used the originally-fitted k_2 for R and F. In addition, we fixed b_2 at –1 for L (for the reason discussed earlier) while looking for the best fitted k_2. The new set of parameters is shown in Table 5. The r^2 values of fitting the model using the parameters in Table 5 are shown in Table 6. Figure 1 displays examples of F_0 curves generated by the model using the parameters in Table 5 (solid lines) together with the original curves (dotted lines).

	Syllable 1			Syllable 2		
	a_1	b_1	k_1	a_2	b_2	k_2
H	—	11.65	.847	—	11.68	.920
R	.66	-15.69	.965	.85	-37.81	(H).972 (R).974 (L).980 (F).976
L	—	3.07	.916	—	-1	.955
F	-.46	23.14	.941	-1.4	78.78	(H).975 (R).974 (L).970 (F).974

Table 5. Fixed parameters for tones in syllables 1 and 2. Tones preceding R and F in syllable 2 are shown in parentheses.

HH	HR	HL	HF	RH	RR	RL	RF
.928	.958	.966	.943	.961	.916	.927	.980
LH	LR	LL	LF	FH	FR	FL	FF
.987	.971	.914	.969	.959	.945	.990	.924

Table 6. r^2 values for fitting the model with parameters in Table 5.

3. DISCUSSION AND CONCLUSION

In this study, we have attempted to quantify the framework proposed by Xu and Wang [4] using a pitch target approximation model. The model inherits the basic assumptions of the framework: a) that F_0 contours in Mandarin are the outcome of implementing pitch targets associated with lexical tones under various articulatory constraints, b) that the pitch targets are either static or dynamic, c) that each pitch target is assigned to and implemented synchronously with a syllable, and d) that the approximation of the pitch target is continuous and asymptotic. Our model further assumes that F_0 contours are generated by the aggregated force of the laryngeal controls. The result of such aggregated force is the movement of F_0 toward a target, which, based on previous observation [1, 4], is asymptotic. Our model thus expresses F_0 contours by an exponential decay function that approximates tonal targets asymptotically. The model also assumes that F_0 movement at the syllable boundary is smooth as long as voicing continues. A pair of hyperbolic tangent functions is used to connect the F_0 curves of two successive tones.

The parameters of the model specify the properties of the pitch target, including its height (b) and direction of movement (a), the rate at which the target is approximated (k), and how smoothly two adjacent tones are joined at the syllable boundary (m).

In our preliminary effort to fit the model to raw data (time-normalized mean F_0 curves of bitonal sequences), we found that the fit of the model was very good when the parameters were allowed to vary. We then tried to fix the parameters by using the same values for different syllable positions for each tone. Fixing the value of m at its grand average resulted in little deterioration of the goodness of fit. Fixing other parameters at their grand averages, however, resulted in greater deterioration. Firstly, we had to use two sets of parameters for all the tonal targets, one for each of the two syllable positions, in order to maintain a reasonably good fit of the model. This suggests that syllable position in Mandarin may introduce additional factors that the model needs to take into consideration. Conceivably, the effects of these factors may be expressed as modifications of either the pitch ranges for the tonal targets or the rate of their approximation, or both.

Secondly, we found that the value of k was rather sensitive to both syllable position and tone. In particular, its value for the dynamic tones (R, F) in syllable 2 was sensitive to the tone in syllable 1. From Table 5, it seems that the greater the difference between $y2_0$ and b, the faster the rate of decay (i.e., a smaller k). It is possible that speakers spend greater effort when $y2_0 - b$ is larger. It is also possible that speakers produce a longer syllable when $y2_0 - b$ is larger. Since the mean F_0 curves in the raw data were time-normalized, longer curves would have been more compressed in time, thus resulting in sharper slopes. Fitting the model to non-time-normalized curves may therefore improve the goodness of fit. This possibility will be tested in our future research.

To conclude, the pitch target approximation model proposed in this study generated F_0 contours that fit closely to the bitonal F_0 contours in Mandarin with variable parameters. We were also able to fix the parameter values across some of the tones and syllable positions. While certain aspects of the model need to be further explored and modified, the basic assumptions of the model seem to have been preliminarily confirmed by the overall consistency of the parameters.

REFERENCES

[1] Xu, Y. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25, 61-83.

[2] Xu, Y. 1998. Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica*, 55, 179-203.

[3] Xu, Y. in press. Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*.

[4] Xu, Y., and Wang, Q. E. in press. Pitch targets and their realization: Evidence from Mandarin. *Speech Communication.*

[5] Zemlin, W. R. 1988. *Speech and Hearing Science — Anatomy and Physiology*. Englewood Cliffs, New Jersey: Prentice Hall.

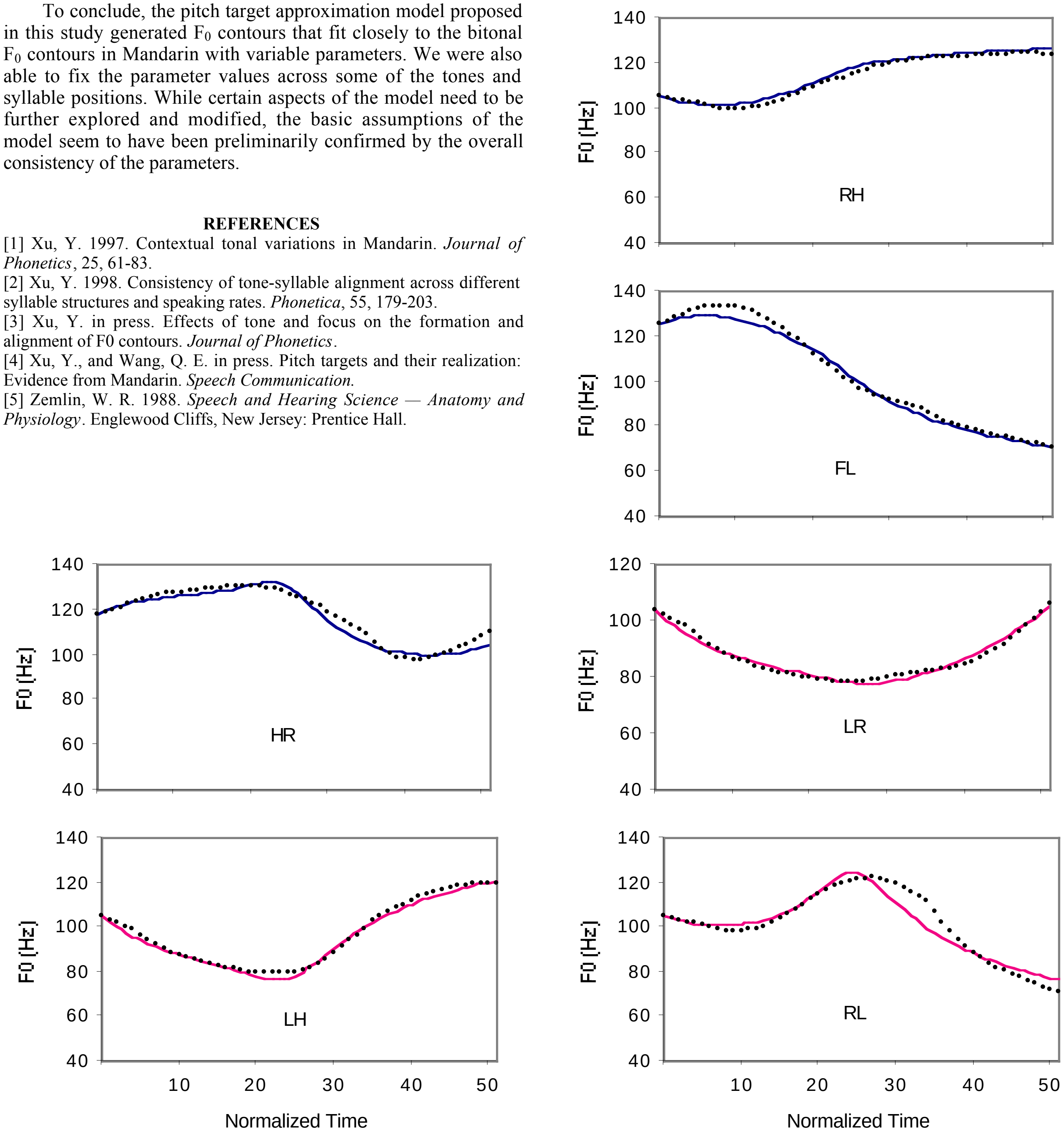

Figure 1. Examples of curve fitting using parameters in Table 5. The solid lines are generated by the model. The RL sequence has the lowest r^2.