FUNCTIONAL MODELING OF TONE, FOCUS AND SENTENCE TYPE IN MANDARIN CHINESE

Santitham Prom-on[a,b], *Fang Liu*[c], *Yi Xu*[a]

[a]Department of Speech, Hearing and Phonetic Science, University College London, UK;
[b]Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand;
[c]Center for the Study of Language and Information, Stanford University, USA
santitham.prom-on@ucl.ac.uk; fliu4@stanford.edu; yi.xu@ucl.ac.uk

ABSTRACT

This paper presents the results of a study on the modeling of the interactions between tone, focus, and sentence type in Mandarin Chinese. We applied the quantitative Target Approximation model (qTA) to an experimental corpus by extracting function-specific qTA parameters and using them to simulate surface F_0 contours. The results demonstrate the accuracy and effectiveness of the method, suggesting that it is a significant step forward toward effective predictive modeling and synthesis of expressive prosody.

Keywords: functional modeling, qTA, tone, focus, sentence type

1. INTRODUCTION

Decades of empirical research has led to a steady accumulation in the understanding of tone and intonation in Mandarin [3-7, 9, 11, 14-16]. In particular, recent findings suggest that F_0 contours of spoken Mandarin result from the interaction of various communicative functions, such as lexical tone, focus and sentence type, both with each other and with the underlying articulatory mechanisms [2, 4-5, 12]. This knowledge makes it possible to computationally generate detailed F_0 contours that closely match those of natural speech. An initial effort has been made in a recent study, showing that F_0 contours of declarative sentences in Mandarin can be generated with an articulatory-functional model that directly controls tone and focus [10]. The present study extends this effort by simulating two aspects of Mandarin prosody that have not been systematically modeled before, namely, the neutral tone and question intonation.

The neutral tone has traditionally been considered as toneless, i.e., having no pitch target of its own because its F_0 varies extensively with the preceding tone [1]. In [2], however, it is argued that the neutral tone does have its own pitch target, except that the target is produced with a weak articulatory force. While this hypothesis has received further support from [6], it has yet to be tested by computational modeling.

Question intonation in Mandarin has been investigated by various studies, including some that involve modeling [3-7, 9, 11, 14-16]. More recently, evidence is shown that sentence type is an independent intonational function whose manifestation is achieved through the articulatory implementation of local tonal pitch target specified by the lexical tones, including the neutral tone [5, 6]. Again, these empirical findings have yet to be confirmed by modeling simulation.

1.1. Functional modeling and analysis-by-synthesis

From a modeling perspective, a model is of little use if it is not *predictive*. To make a model predictive, however, it is critical to first determine what the predictors should be. If, as suggested above, communicative functions like tone, focus and sentence type and their interactions are directly behind the complex surface F_0 contours in Mandarin, these communicative functions should then be the predictors. An alternative to such *functional modeling* is to simulate F_0 with predictors whose functional status is ambiguous, or whose definition includes characteristics of observed F_0 patterns, e.g., pitch accents, F_0 turning points, etc.

From a theoretical perspective, functional modeling provides a powerful tool for hypothesis testing. That is, by assessing how well surface F_0 contours generated based on a set of hypothesized predictors, investigators can validate or falsify both general and specific theoretical assumptions about tone and intonation. Such a process is known as *analysis-by-synthesis*.

1.2. qTA model

The computational model used in the present study is the quantitative Target Approximation (qTA) model. This model simulates the production of tone

and intonation as a process of syllable-synchronized sequential target approximation [10, 14]. Fig. 1 illustrates the basic idea of target approximation [14]. The qTA model represents F_0 as the surface response of the target approximation process which is driven by pitch targets. A pitch target is a forcing function representing the joint force of the laryngeal muscles that control vocal fold tension. It is represented by a simple linear equation,

$$x(t) = mt + b \quad (1)$$

where m and b denote the slope and height of the pitch target, respectively.

The F_0 control is implemented through a third-order critically damped linear system, in which the total response is

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

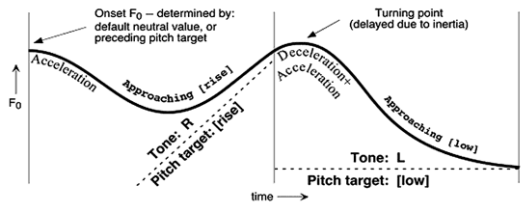
where the first term $x(t)$ is the forced response of the system which is the pitch target and the second term is the natural response of the system. The transient coefficients c_1, c_2 and c_3 are calculated based on the initial F_0 dynamic state and the pitch target of the specified segment. The parameter λ represents the strength of the target approximation movement. In qTA, the initial F_0 dynamic state consists of initial F_0 level, $f_0(0)$, velocity $f_0'(0)$, and acceleration, $f_0''(0)$. The dynamic state is transferred from one syllable to the next at the syllable boundary to ensure continuity of F_0. The three transient coefficients are computed with the following formulae.

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

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

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

Figure 1: The Target Approximation process [14].



2. METHODS

2.1. Corpus

The corpus was originally designed for a study to test the effects of sentence type, focus and tone on the F_0 contours of the full tones (i.e., Tones 1-4) as well as the neutral tones in Mandarin and their interaction with each other [6]. The corpus was recorded by 8 native Mandarin speakers. Each target sentence consists of 8 syllables. The tone of the third syllable varies across all the full tones, including High (H, T1), Rising (R, T2), Low (L, T3) and Falling (F, T4). The first syllable is always H and the second syllable always L. The fourth to sixth syllables are always the Neutral tone (N, T0). The final two syllables are either both H or both N. Each sentence was also said as either a statement or a question, and with focus either on the second or third syllable. Thus there are 32 combinations in total. For each combination, the utterance was repeated five times by each speaker. The structure of the corpus is shown in Table 1.

Table 1: Sentence structure of the corpus.

<table>
<tr><td rowspan="4">ta1 mai3
H L
他买</td><td>ma1 ma0
H N
妈妈</td><td rowspan="4">men0 de0
N N
们的</td><td rowspan="2">le0 ma0
N N
了嘛</td></tr>
<tr><td>ye2 ye0
R N
爷爷</td></tr>
<tr><td>nai3 nai0
L N
奶奶</td><td rowspan="2">mao1 mi1
H H
猫咪</td></tr>
<tr><td>mei4 mei
F N
妹妹</td></tr>
</table>

2.2. Functional parameter estimation

For each syllable in an utterance, the qTA parameters m, b and λ are simultaneously estimated by iteratively searching for the optimal parameter set with the lowest sum square error. The search space for each parameter can be optionally limited by a symbolic label assigned to the syllable. For example, if the syllable is labeled as R (Rising), m is searched only from zero to a predefined maximum value. In the present study, the qTA parameter estimation was done with a modified version of PENTAtrainer [13].

The parameter estimation for the neutral tone required special treatment. To derive the hypothesized weak articulatory strength [2], we grouped consecutive neutral tone syllables together as one chunk during initial parameter estimation. The strategy is based on the observation that weak articulatory strength would lead to slow approximation of the neutral target, which typically takes several consecutive neutral tone syllables [2]. Grouping them together to obtain a single target means that the assessment algorithm would not mistakenly assume full target achievement within individual neutral tone syllables.

After obtaining the optimal parameter set for each syllable in all utterances, the qTA parameters were clustered into categories by tone, focus and

sentence type. For each category in a particular combination of functions, a parameter template, consisting of *m*, *b*, λ and syllable duration (*d*), was summarized by calculating the median of each parameter. The median was used instead of the mean because of its robustness against outliers. The templates were estimated either from all speakers, for speaker independent testing, or from each specific speaker, for speaker dependent testing.

3. RESULTS

3.1. Accuracy of synthesis

Table 2 shows means and standard errors of root mean square error (RMSE) and Pearson's correlation coefficient for various functional combinations. The low RMSE and high correlation for the *Resynthesis* condition indicate that the qTA model can simulate the F_0 of the original utterances quite well. The high RMSE and low correlation in the Tone only condition suggest the inadequacy of simulating F_0 with tone alone. Adding Focus and Sentence Type as additional control functions improves the synthesis accuracy both in terms of RMSE and correlation. Additionally, the results are not significantly different between the speaker-dependent and -independent conditions. This seems to demonstrate the generalizability of the functional modeling approach.

Table 2: Average RMSE in semitone and correlation coefficients of each imposed function or interaction. Parameters of each function are averaged either within (as speaker dependent cases) or across speakers (as speaker independent cases). The number of parameter indicates the total number of parameter set used in F_0 synthesis.

Imposed Function	Speaker Dependent			Speaker Independent		
	Number of parameter	RMSE (st)	Correlation	Number of parameter	RMSE (st)	Correlation
Resynthesis	10240	0.84 ±0.08	0.98 ±0.00	-	-	-
Tone	64	4.35 ±0.36	0.69 ±0.02	8	4.55 ±0.28	0.67 ±0.02
Tone + Focus	120	4.05 ±0.31	0.75 ±0.01	15	4.38 ±0.28	0.72 ±0.03
Tone + Sentence Type	128	3.73 ±0.31	0.73 ±0.02	16	4.01 ±0.27	0.70 ±0.02
Tone + Focus + Sentence Type	240	3.37 ±0.26	0.79 ±0.01	30	3.78 ±0.27	0.75 ±0.03

* standard errors are calculated across all speakers

3.2. Further analysis

Table 2 also indicates that there are certain interactions between focus and sentence type functions. This can be seen in the fact that imposing either focus or sentence type function yielded different improvements. But when both were included, the improvement was more than the sum of the effects of both functions. Such interaction was thus captured by the categorical representation of the communicative function.

Fig. 2 shows selected examples of F_0 contours generated from speaker independent parameters when tone, focus and sentence type are alternatively included as predictors. As more functions are added, the synthetic F_0 contours become increasingly closer to the natural ones. Note in particular that the parameters used were obtained from all speakers, but the natural contour is from a single speaker.

Tables 3-5 show the functional parameters for statement and question in the on-focus, post-focus and final word, respectively. Interestingly, the differences in the on-focus parameters of full tones between statement and question are quite subtle (Tables 3-4). They are apparently sufficient to generate F_0 contours that closely match the natural ones, as indicated by Table 2. The post-focus *b* for the neutral tone in statement is lower than that in question, while its λ is low in both statement and question. This low strength seems to be consistent with the weak strength hypothesis for the neutral tone [2]. Table 5 shows that target heights of final words are much lower in statement than in question, which is consistent with previous finding of accelerated F_0 contrast between statement and question toward the end of the sentence [5].

Figure 2: Examples of F_0 contours synthesized using different functional combinations; A) tone only, B) tone + focus, C) tone + sentence type, D) tone + focus + sentence type. (solid: synthesized, dashed: natural).

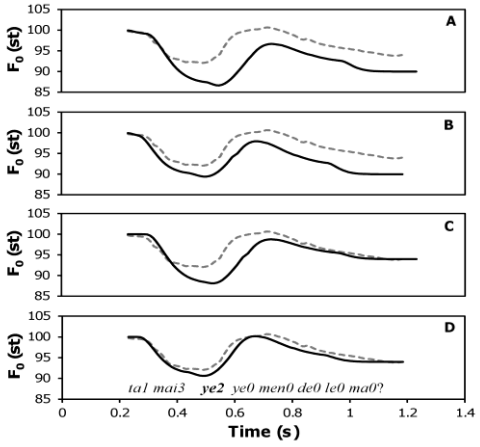


Table 3: Averaged on-focus parameters comparing between (S)tatement and (Q)uestion.

Tone	*m* (st/s)		*b* (st)		*λ*		*d* (s)	
	S	Q	S	Q	S	Q	S	Q
H	0	0	2	3	39	42	.203	.198
R	110	115	-8	-5	30	33	.175	.160
L	0	0	-16	-14	29	30	.260	.241
L-S	119	113	-2	-2	30	30	.279	.274
F	-3	-1	4	4	53	58	.190	.189

* L-S: Low tone (Sandhi)

Table 4: Averaged post-focus parameters comparing between (S)tatement and (Q)uestion.

Tone	*m* (st/s)		*b* (st)		*λ*		*d* (s)	
	S	Q	S	Q	S	Q	S	Q
N	0	0	-11	-5	20	19	.106	.104
H	0	0	-3	-2	65	63	.142	.148
R	61	79	-8	-6	81	68	.140	.134
L	0	0	-14	-10	27	29	.154	.151
F	-4	-3	-2	-1	79	70	.139	.143

Table 5: Averaged pitch target parameters of syllables in a final word comparing between (S)tatement and (Q)uestion.

Tone	*m* (st/s)		*b* (st)		*λ*		*d* (s)	
	S	Q	S	Q	S	Q	S	Q
N	0	0	-14	-6	58	48	.132	.143
H	0	0	-11	-3	56	53	.183	.189

4. CONCLUSION

This study has demonstrated the effectiveness of quantitative modeling based on the qTA model, and that of the analysis-by-synthesis strategy to simulate interactions between tone, focus and sentence type. Instead of using models to only analyze changes in parameters as previously done [8-9], this paper takes on another crucial step by systematically simulating the actual prosodic phenomena. The modeling data such as F_0 contour comparisons, statistical error reports and parameter distributions in the study provide insights into the interaction between tone, focus and sentence type. This study also, for the first time, successfully simulated the neutral tone based on the concept of weak articulatory strength accompanying a [mid] target [2].

5. ACKNOWLEDGEMENTS

The authors would like to thank the Royal Society and the Royal Academy of Engineering for the financial support through the Newton International Fellowship Scheme, and the Thai Research Fund for the TRF-CHE Research Grant for New Scholar with grant number MRG5380038. This work was supported in part by the Economic and Social Research Council with grant number PTA-026-27-2480 to F.L.

6. REFERENCES

[1] Chao, Y.R. 1968. *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Press.

[2] Chen, Y., Xu, Y. 2006. Production of weak elements in speech – Evidence from f0 patterns of neutral tone in standard Chinese. *Phonetica* 63, 47-75.

[3] Ho, A.T. 1977. Intonation variation in a Mandarin sentence for three expressions: interrogative, exclamatory and declarative. *Phonetica* 34, 446-457.

[4] Lin, M. 2004. On production and perception of boundary tone in Chinese intonation. *Proc. TAL 2004*, 125-129.

[5] Liu, F., Xu. Y. 2005. Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62, 70-87.

[6] Liu, F., Xu, Y. 2007. The neutral tone in question intonation in Mandarin. *Proc. Interspeech 2007*, 630-633.

[7] Ma, J.K.Y., Ciocca, V., Whitehill, T.L. 2006. Quantitative analysis of intonation patterns in statements and questions in Cantonese. *Proc. Speech Prosody 2006*, 277-280.

[8] Mixdorff, H., Pfitzinger, H.R. 2009. A quantitative study of F0 peak alignment and sentence modality. *Proc. Interspeech 2009*, 1003-1006.

[9] Ni, J.F., Kawai, H. 2004. Pitch target anchor Chinese tone and intonation patterns. *Proc. Speech Prosody 2004*, 92-98.

[10] Prom-on, S., Xu, Y., Thipakorn, B. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *J. Acoust. Soc. Am.* 125, 405-424.

[11] Shen, X.N.S. 1990. *The Prosody of Mandarin Chinese*. Berkeley, CA: University of California Press.

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

[13] Xu, Y., Prom-on, S., 2010-2011. PENTAtrainer.praat. *http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer/*

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

[15] Yuan, J. 2004. *Intonation in Mandarin Chinese: Acoustics, Perception, and Computational Modelling.* Ph.D. dissertation, Cornell University.

[16] Zheng, X.L., Martin, P., Boulakia, G. 2004. Tone and intonation in declarative and interrogative sentences in Mandarin. *Proc. TAL 2004*, 235-238.