

Emotional Intonation Modeling: a Cross-Language Study on Chinese and Japanese

Ai-Jun Li*, Yuan Jia*, Qiang Fang* and Jian-Wu Dang†

* Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China.

E-mail: liaj@cass.org.cn. Tel: +86-10-65237408

† Tianjin University, Tianjin, China;

‡ Japan Advanced Institute of Science and Technology, Japan

E-mail: jdang@jaist.ac.jp

Abstract—This study attempts to apply PENTA model to simulate the emotional intonations of two typologically distinct languages, the tone language of Mandarin Chinese and the pitch accent language of Japanese. First, the overall F0 features of the emotional intonations of 4 speakers were analyzed and contrasted across seven emotions and across two languages. And then the performances of the qTA model for simulating each language were numerically evaluated and compared within and across the two languages. The results showed that F0 features have bigger distinctions across the two languages than within them. The qTA model can efficiently encode emotional or pragmatic information for both Chinese and Japanese.

I. INTRODUCTION

As far as the present study is concerned, we mainly focus on emotional prosody, or communicative prosody, the term proposed by Sagisaka [14,15], which is more related to emotion/affect simulation or synthesis. A number of models have been proposed for intonation/prosody generation, such as Fujisaki Model [2], SFC [1], STEM-ML model [4], the PENTA model/qTA [12-13, 16-18] and Daniel Hirst's Momel and Intsint algorithms [3] for the automatic analysis of speech prosody. Xu [19] gave a critical review of methodology and progress in speech prosody. Apart from Fujisaki's model, most of these models have not been applied to the simulation of emotional speech.

PENTA/qTA model (Parallel Encoding and Target Approximation model), proposed by Xu, is based on the assumption that the surface prosody represents both the articulatory mechanisms and communicative functions, while the communicative functions control F0 contours via specific (and parallel) encoding schemes. [133, 184] It has been widely evaluated and applied in several languages [11] by checking the encoding schemes including the neutral tone, focus, sentence types (interrogation / statement), the coding and speaker styles (read/spontaneous) and the target implementation domain (syllable/word) [11, 13, 16].

While, this study extends the domain of the PENTA model from neutral intonation to emotional intonation in two typologically different languages, the tone language of Mandarin Chinese and the pitch accent language of Japanese.

First, the overall F0 features of the emotional intonations of four speakers are analyzed and contrasted across seven

emotions and across two languages. Then the performances of the qTA model in simulating each language are compared objectively or numerically within each language for two speakers and between the two languages as well. Finally, the performances of qTA in simulating these two emotional intonations will be compared with those of the neutral intonations for other languages including Mandarin, English, Brazilian Portuguese and German.

II. THE PENTA AND THE qTA MODELS

The Parallel Encoding and Target Approximation (PENTA) model was proposed by Xu based on the assumption that the surface prosody represent both the articulatory mechanisms and communicative functions, the communicative functions control F0 contours via specific (and parallel) encoding schemes [13]

Xu and Prom-on made a general review on the present prosodic models and described the principle of PENTA and the quantitative target approximation (qTA) model in [12,13] as shown in Fig.1 and Fig.2.

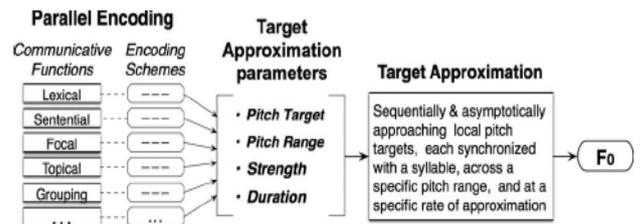


Fig.1 A sketch of the PENTA model from[12,13]

qTA has been tested numerically to simulate tone, lexical stress and focus in Mandarin and English with an automatic analysis-by-synthesis procedure [8]. The core of qTA is the target approximation mechanism. In qTA, a pitch target is a forcing function representing the joint muscular force of the laryngeal muscles that control vocal fold tension. It is represented by a simple linear equation, as in (1)

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

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

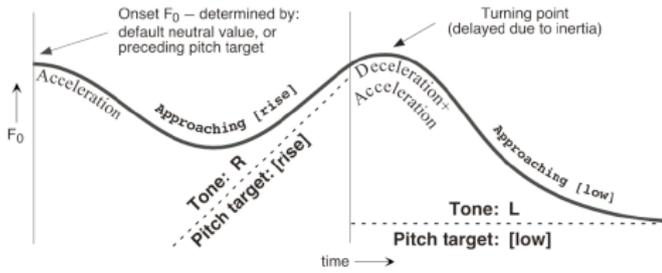


Fig.2. Illustration of the TA model.

The vertical lines represent syllable boundaries. The dashed lines represent underlying pitch targets. The thick curve represents the F0 contour that results from asymptotic approximation of the pitch targets [12,13].

The control of the vocal fold tension in qTA is implemented through a third-order critically damped linear system, in which the total response is

$$F0(t) = x(t) + (c1 + c2.t + c3.t^2).e^{-\lambda.t} \quad (2)$$

where $x(t)$ is the forced response of the system which is the pitch target and the second part is the natural response, whereas the λ parameter specifies the rate of approaching the target line. The coefficients $c1$, $c2$ and $c3$ are related to the initial conditions. Thus, given these initial conditions, three parameters completely specify the modeled F0 contour for a given linguistic segment. Thus, the qTA model generates a left-to-right local implementation of a F0 contour where the absolute F0 values, together with the first and second F0 derivatives in an utterance segment are defined as the last values of the immediately previous utterance segment.

Since the encoding schemes are hypothesized to be language-specific, this study tries to explore the encoding schemes relating emotional intonation for two typologically different languages: Chinese and Japanese.

For emotional evaluation, i.e. Comparisons of qTA generated F0 contours with those of natural speech are showed in terms of RMSE, Correlation, and perceptual judgment of naturalness [12,13]. In the present paper, the perceptual judgment is excluded.

III. CHINESE AND JAPANESE EMOTION DATA

The Chinese data employed in the study were obtained from the emotional speech corpus Emotion CASS [5-7]. The text material includes a set of 111 sentences with various length (1~14 syllables), different types (narrative/interrogative), syntactic structures and tonal combinations. The disyllabic sentences contain sixteen tonal combinations.

The Japanese emotion data were obtained from our cross-culture study on emotional speech [9]. The text material includes 10 sentences with various sentence length (1~5 syllables), tonal combinations and grammatical structures.

Both Chinese and Japanese texts are expressive neutral. Two Chinese speakers (male SONG & female AN) and two Japanese speakers (male ZD & female YD) were recruited. The Chinese are university students from a professional oral

dubbing department. The Japanese are international students from Tokyo studying in Beijing University. They were selected as the speaker because they are skillful in oral expression.

Seven emotions including ‘Disgust, Sad, Angry, Happy, Surprise, Fear and Neutral’ were recorded in our sound proof room, resulting 111*7*2=1554 Chinese emotional utterances and 10*7*2=140 Japanese emotional utterances.

All the emotional utterances were annotated with syllable boundaries, and F0 data were extracted and manually corrected by using Praat [10]. The annotated data were stored in TextGrid as requested by qTA trainer.

IV. MODELING THE EMOTIONAL INTONATION USING QTA

The PENTA trainer version 1.4 script [20], run under Praat, was used for modeling the F0 contours. All utterances were segmented into syllables, within which the search for the three qTA parameters, target height and slope, and target strength, took place.

In qTA, two measurements were used to assess the degree of similarity of fit between synthesized and natural F0, RMSE and Pearson’s Correlation coefficient. RMSE measures the difference between natural and synthesized F0 contours and Correlation indicates the linear relationship between them. Correlation is used for evaluating F0 contours, not the model parameters. Positive high Correlation indicates consistency between original and synthesized F0 contours not only in height but also in contour shapes. The semitone scale is used for measuring RMSE so as to assess the performance across different speakers.

A. Chinese emotional intonation simulating

The PENTA trainer version 1.4 script, running under Praat, is used for modeling the F0 contours as shown in Fig.3 where the simulated F0 curve of this ‘Happy’ utterance is well fitted.

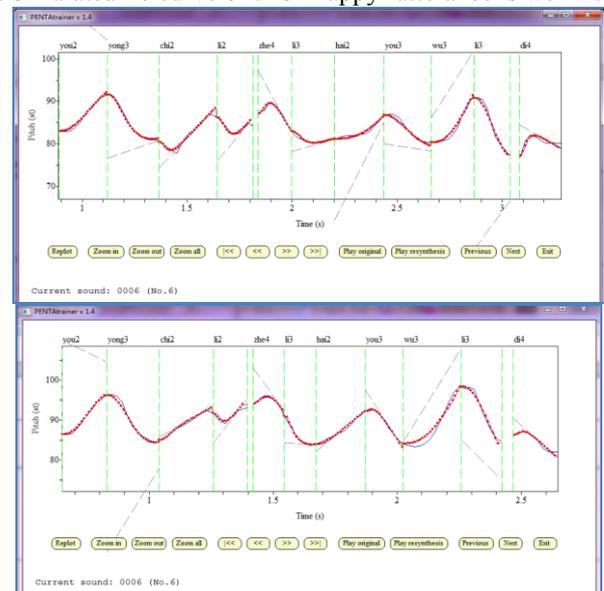


Fig.3 Two synthesized Chinese utterance ‘游泳池离这里还有5里地 (The swimming pool is 5 miles away from here.)’ by qTA: the upper panel is for ‘Neutral’ emotion and the lower panel is for ‘Happy’ emotion. The blue lines

are the original intonations and the red dotted lines are the simulated intonations.

a. F0 features of Chinese emotional intonation

Figures 4 and 5 depict the mean F0 features of minF0, maxF0 and meanF0 in seven emotions for the two Chinese. The figures show that ‘Happy, Surprise and Angry’ have higher F0 than other emotions for both speakers.

Comparing the two figures, we see that the variation patterns across seven emotions are consistent except that speaker AN has a lower F0 in ‘Sad’ intonation than speaker SONG.

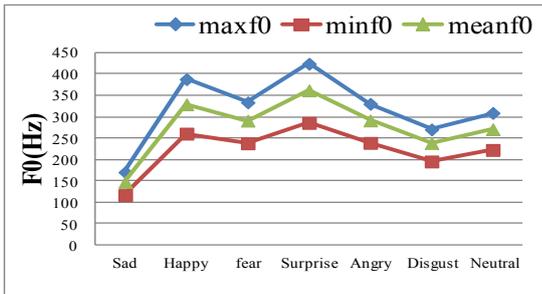


Fig. 4 F0 variations of minF0, maxF0 and meanf0 in seven emotions for Chinese female speaker AN.

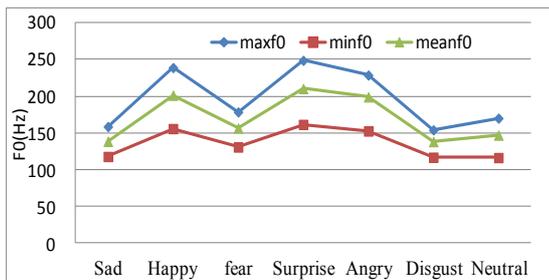


Fig. 5 F0 variations of minF0, maxF0 and meanf0 in seven emotions for Chinese male speaker SONG.

The mean F0 excursion sizes are bigger for ‘Happy and Surprise’ and smaller for ‘Fear and Disgust’ as shown in Fig.6. But speaker AN has bigger size for ‘Sad’ and speaker SONG has smaller size for ‘Sad’.

ANOVA analysis on minF0, maxF0 and F0 excursion size indicated that these features are all significantly different across seven emotions for both speakers. (speaker An: maxf0: $F(6,3311)=942.30, P=0.0$; minf0: $F(6,3311)=550.38, P=0.0$; meanf0: $F(6,3311)=819.37, P=0.0$; excursion size: $F(6,3311)=13.20, P=0.0$). Speaker SONG: maxf0: $F(6,3314)=503.05, P=0.0$; minf0: $F(6,3314)=182.66, P=0.0$; meanf0: $F(6,3314)=353.04, P=0.0$; excursion size: $F(6,3314)=57.91, P=0.0$).

The mean F0 excursion sizes are 6.3St and 6.4St for speakers AN and SONG respectively. The F0 features demonstrate that Chinese speakers may adopt different strategies to express emotions.

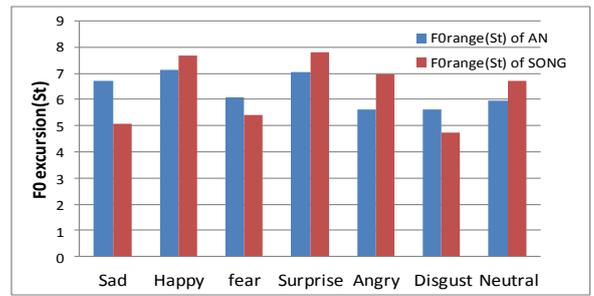


Fig.6 The mean F0 excursion size of the two Chinese speakers in seven emotions.

b. qTA performance of Chinese emotional intonations

How about performance of the qTA to simulate the Chinese intonations in different emotions with so many acoustic deviations? We adopted qTA to simulate 1554 utterances and got the evaluation coefficients of RESE and Correlations. The following figures 7 and 8 show PTA coefficients of mean RMSE and mean Correlations for the two Chinese speakers.

The mean correlations (Standard deviations) are 0.75 (0.02) and 0.76(0.11), the mean RMSEs (Standard deviations) are 0.58St (0.03) and 0.57St(0.09) for speaker AN and speaker SONG respectively.

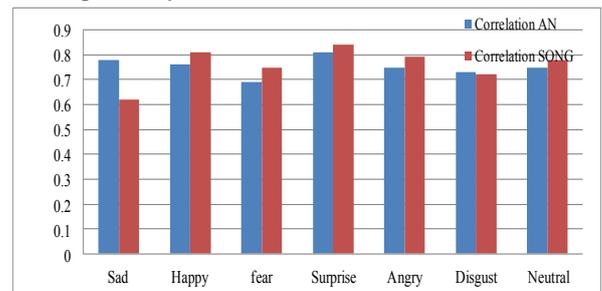


Fig. 7 Mean Correlations of seven emotions for two Chinese speakers.

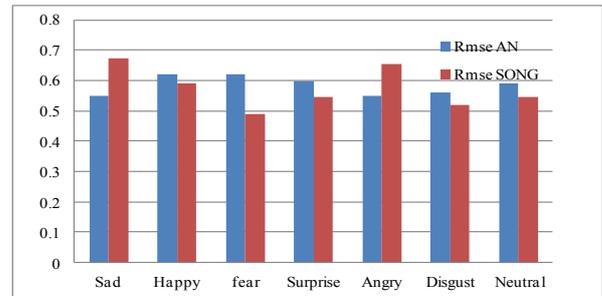


Fig. 8 Mean RMSE of seven emotions for two Chinese speakers.

ANOVA analysis showed that Correlations are significantly different across seven emotions for both Chinese speakers (An: $F(6,3311)=9.94, P=0.00<0.05$; SONG: $F(6,3314)=40.10, P=0.00<0.05$) ; RMSEs are significantly different across seven emotions for both Chinese speakers (An: $F(6,3311)=2.32, P=0.03<0.05$; SONG: $F(6,3314)=13.07, P=0.00<0.05$).

TABLE I
ANOVA ANALYSIS FOR RMSE AND CORRELATION BETWEEN TWO CHINESE SPEAKERS

Emotion	P value of RMSE	P value of Corr.
Sad	.000	.000
Happy	.385	.005
fear	.000	.000
Surprise	.040	.040
Angry	.000	.008
Disgust	.097	.476
Neutral	.130	.123

The differences of Correlation and RMSE between the two Chinese speakers were statistically analyzed as shown in Table I. The result indicated that ‘Disgust and Neutral’ have similar ‘Correlations’ while other emotions have different ‘Correlations’ between the two speakers; ‘Disgust Neutral and Happy’ have consistent RMSE while the others have different RMSE between the two speakers.

The results also suggested that qTA performances on ‘Happy and Surprise’ are higher than ‘Neutral’ speech, lower for ‘Fear and Sad’ in some cases. The lower performances may be caused by the ‘abnormal’ F0 contours such as the trill sound in ‘Sad and Fear’.[1]

B. Japanese emotional intonation simulating

Based on the F0 data and syllable annotation of the emotional utterances, the simulated emotional intonations were obtained by running Praat script of PENTA trainer on 140 Japanese emotional utterances. Fig. 9 shows two simulated intonation curves which express ‘Neutral and Angry’ emotions.

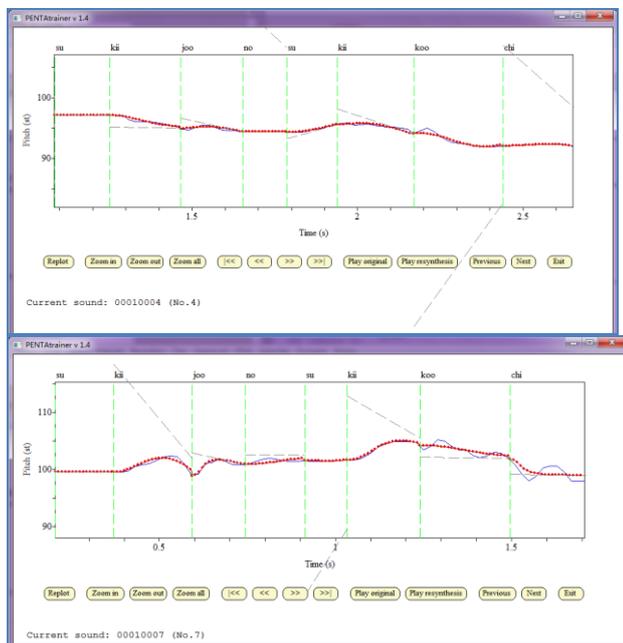


Fig. 9 Two synthesized Japanese utterance ‘スキー場のスキーコーチ (coach of ski resort)’ by qTA: upper panel is for ‘Neutral’ emotion and lower panel is for ‘Angry’ emotion. The blue lines are the original intonations and the red dotted lines are the simulated intonations. (Green lines mark the syllable onsets, grey dotted lines are linear target lines)

a. F0 features of Japanese emotional intonation

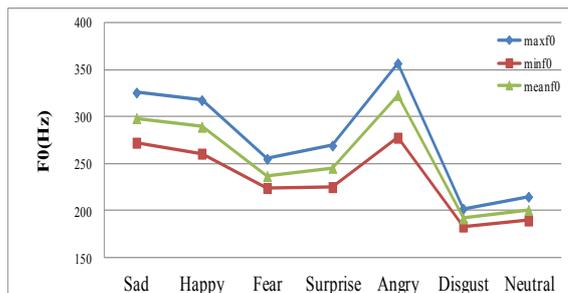


Fig. 10 F0 variations of minF0, maxF0 and meanF0 in seven emotions for Japanese female speaker YD.

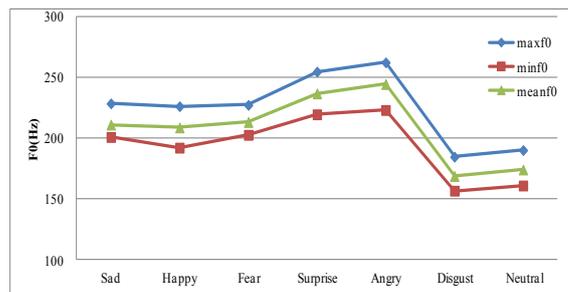


Fig. 11 F0 variations of min F0, max F0 and meanF0 in seven emotions for Japanese male speaker ZD.

F0 features of two Japanese speakers are shown in figures 10 and 11. Based on the ANOVA analysis, it indicated that F0 varies significantly across seven emotions in maxf0, minf0 and meanf0 for both speakers (all $P < 0.05$). For female speaker YD, ‘Happy, Angry and Sad’ have highest F0 while ‘Disgust and Neutral’ have lowest F0. For speaker ZD ‘Angry and Surprise’ have highest F0 while ‘Disgust and Neutral’ have lowest F0.

For speaker YD (Fig. 12), the mean ‘excursion_size’ is 2.93St, but the excursion_sizes are significantly different across seven emotions ($F(6,382) = 6.883, P = 0.0$), in which ‘Angry, Happy, Surprise and Sad’ have wider F0 rang than other emotions. For speaker ZD (Fig. 9), the mean ‘excursion_size’ is 2.64St, which has no significant difference across seven emotions ($F(6,377) = 1.095, P = 0.365$).

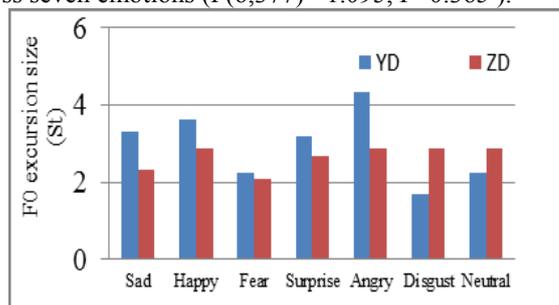


Fig 12: The mean F0 excursion size of the two Japanese speakers in seven emotions.

Comparing the overall F0 varying patterns in Fig. 10 with Fig. 11, we conducted that the two Japanese have less consistent patterns than the two Chinese, especially in ‘Fear and Surprise’ emotions, which suggests that the two Japanese

speakers expressed these two emotions in quite different strategies.

Next section we will focus on how the performance on qTA to simulate the Japanese emotional intonations with so large interpersonal and across emotional deviations.

b. qTA performance of Japanese emotional intonations

The following Figures 13 and 14 show qTA coefficients of mean RMSE and mean Correlation for the two Japanese.

The ‘Correlations’ keep consistent across seven emotions for both two speakers, for YD around 0.71 ($F(6,382)=0.882, P=0.508$); for ZD, around 0.76 ($F(6,377)=0.1461, P=0.191$).

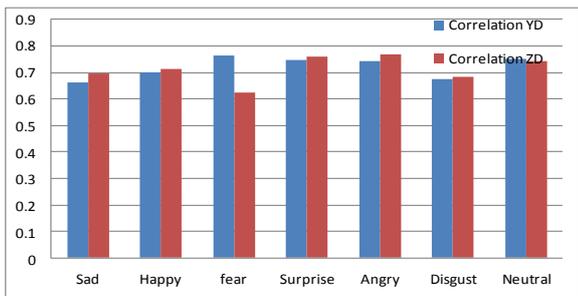


Fig. 13 Mean Correlations of seven emotions for two Japanese speakers.

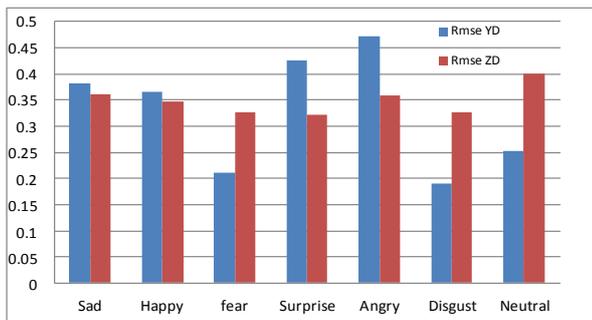


Fig. 14 Mean RMSEs of seven emotions for two Japanese speakers.

For speaker YD, the ‘RMSEs’(mean=0.33) are significantly different across seven emotions ($F(6,382)=6.356, P=0.0$). ‘Disgust Fear and Neutral’ have smaller Rmses than the others. While for ZD, the ‘RMSEs’ (mean=0.35) are not significantly different across seven emotions ($F(6,377)=0.352, P=0.909$).

The results suggested that the qTA performance of emotional speech is as good as ‘Neutral’ speech.

C. Summary on emotional intonations across two languages

a. Overall F0 features

Based on F0 variation patterns for seven emotional states as depicted in Figures 4, 5, 10 and 11, it explicitly showed that Chinese and Japanese express emotions with quite distinct strategies on intonations, but this distinction within the language is smaller than across Languages. For an instance, Chinese always use higher pitch in expressing ‘Happy and Surprise’ but Japanese don’t. Chinese ‘Sad’ has lower F0 while Japanese don’t. Moreover, the mean F0 excursion sizes

of the Japanese are almost 3St smaller than those of the Chinese as shown in Fig.4 and Fig. 10, and were summarized in the 1st column in table II.

b. PTA performance

The four speakers have rather consistent mean Correlations ranging from 0.71-0.76 as listed in Table II. Although the mean RMSEs are similar within languages, they are quite different across two languages. The values of Correlation are slightly higher for Chinese but the values of RMSE are smaller for Japanese, so it is hard to reveal which language has better performance than another.

TABLE II.
MEAN PARAMETERS OF THE SEVEN EMOTIONS FOR TWO LANGUAGES

Speakers	F0 Excursion Size(St)	CORR.	RMSE
Chinese AN	6.3	0.75	0.58
Chinese SONG	6.4	0.76	0.57
Japanese YD	2.93	0.71	0.33
Japanese ZD	2.64	0.76	0.35

V. DISCUSSION AND CONCLUSION

The F0 features reveal that the overall variations of intonations are more distinct between Chinese and Japanese than within the two languages. That could be caused by both the typological language distinctions and the cross-cultural differences. In other words, since Chinese is a tonal language, its F0 contour should convey both tone and intonation information, so that the F0 excursion size of Chinese are bigger than that of Japanese. Another reason might be that the Chinese speakers are a professional actor and actress, who vocally expressed the emotions in a quite exaggerated way., while the Japanese speakers expressed the emotions more by using facial actions than vocal actions [8].

TABLE III
CORRELATION AND RMSE COEFFICIENTS IN SIMULATING SEVERAL LANGUAGES.

Language and style	CORR.	RMSE
Read Chinese with tona/focus/position[16]	0.76	2.42
Emotional Chinese (present study)	0.76	0.58
Emotional Japanese (present study)	0.74	0.34
Read English(stress/position)[16]	0.77	1.72
Read BP (Barbosa) [11]	0.99	1.0
Storytelling BP[11]	0.99	1.2
Read German[11]	0.90	1.4
Storytelling German[11]	0.92	1.3

To check the qTA performances of the simulated emotional intonations for the two languages, the numeric assessment was made on the Correlation and RMSE values, the results indicated that the qTA is efficient for modeling both Japanese and Chinese emotional intonations. Table III summaries these two coefficients when simulating several languages of Chinese [17,18], English[13,16], Japanese, German and Brazilian Portuguese (BP) [11] in different contexts. It reveals that qTA model have quite similar performance in encoding

the Chinese and Japanese emotional speech as encoding focus and tone contexts in English and Chinese whose pitch targets are implemented in Syllable domain, while German and BP (Brazil Portuguese) have higher performance than Chinese, English and Japanese. One of the reasons for this is that tonal targets of German and BP are implemented in word domain in [11].

Until now, we may conclude that PENTA model can be used for modeling multi-language intonations including tone language, pitch accent language and stress language.

Besides the numerical assessment, the perceptual assessment has also been made, which indicated that some new encoding scheme has to be adopted for Chinese emotional intonations to express the attitudinal/pragmatic functions as a communicative function.[9]

In previous study on Chinese speech, we know that prosodic features and intonation patterns are quite different between read and spontaneous speech.[21] It seems that the prosodic features of Chinese spontaneous speech are closer to the stress languages as English. So it would be interesting to use PENTA model to typologically check the F0 fitting performance of qTA between spontaneous and read speech with varying target domains across different languages. This is the work we are carrying on.

ACKNOWLEDGMENTS

Thanks Dr. Yi XU for providing the script and discussing with the author. This work was supported by the National Basic Research Program (973 Program) of China (No. 2013CB329301), NSFC Project with No. 61233009 and 60975081 and CASS innovation project.

REFERENCES

- [1] G. Bailly, and B. Holm, "SFC: a trainable prosodic mode," *Speech Communication*, vol. 46, pp. 348-364, 2005.
- [2] H. Fujisaki, and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan*, no. 4, pp. 233-242, May 1984.
- [3] D. J. Hirst, and A. D. Cristo, "Intonation Systems, A survey of Twenty Language," *Cambridge University Press*, 1998.
- [4] G. Kochanski, and C. Shih, "Prosody modeling with soft templates," *Speech Communication*, vol. 39, pp. 311-352, 2003.
- [5] A. J. Li, Q. Fang, and J. W. Dang, "Emotional intonation in a tone language: experimental evidence from Chinese," *ICPhS'2011*.
- [6] A. J. Li, Q. Fang, and J. W. Dang, "Emotional Expressiveness of Successive Addition Boundary Tone in Mandarin Chinese," *Speech Prosody, Shanghai, China*, 2012.
- [7] A. J. Li, Q. Fang, Y. Jia, and J. W. Dang, "Successive Addition Boundary Tone in Chinese Disgust Intonation," *NACCL24, USA*, June, 2012.
- [8] A. J. Li, and J. W. Dang, "A Cross-Cultural Investigation on Emotion Expression under Vocal and Facial Conflict - Also an observation on Emotional McGurk Effect," *International symposium on biomechanical and physiological modeling and speech science, Kanazawa, Japan*, 2009.
- [9] A. J. Li, Q. Fang, Y. Jia, and J. W. Dang, "More Targets Simulating Emotional Intonation of Mandarin with PENTA," *ISCSLP'2012*, pp. 271-275, 2012.
- [10] P. Boersma, and D. Weenink, "Praat," *an open resource on line: <http://www.fon.hum.uva.nl/praat/>*.
- [11] A. Plínio, Barbosa, H. Mixdorff, and S. Madureira, "Applying the quantitative target approximation model (qTA) to German and Brazilian Portuguese," *Interspeech2011, Italy*, pp. 2025-2028, 2011.
- [12] S. Prom-on, and Y. Xu, "Articulatory-Functional Modeling of Speech Prosody: A Review Proc." *Interspeech2010, Makuhari*, pp. 46-49, 2010.
- [13] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *J. Acoust. Soc. Am.*, vol. 125, no. 1, pp. 405-424, 2009.
- [14] Y. Sagisaka, "Modeling prosody variations for communicative speech and the second language towards trans-disciplinary scientific understanding," *Keynote speech of Speech Prosody*, 2012.
- [15] Y. Sagisaka, et al., "Prosody generation for communicative speech synthesis," *SNLP'2005*, pp. 23-28.
- [16] Y. Xu, and C. H. Xu, "Phonetic realization of focus in English declarative intonation," *J. Phon.*, vol. 33, pp. 159-197, 2005.
- [17] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, pp. 220-251, 2005.
- [18] Y. Xu, and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication*, vol. 33, pp. 319-337, 2001.
- [19] Y. Xu, "Speech prosody: a methodological review," *Journal of Speech Sciences*, no. 1, pp. 85-115, 2011.
- [20] Y. Xu, and S. Prom-on, "PENTA Trainer," *praat, <http://www.phon.ucl.ac.uk/home/yi/PENTATrainer/>*.
- [21] Y. B. Liu, and A. J. Li, "The difference between read and spontaneous spoken Chinese," *J. of Chinese Information Processing*, vol. 16, no. 1, pp. 13-18, 2002.