Naturalness Judgement of L2 Mandarin Chinese - Does timing matter?

Chiharu Tsurutani[1], *Dean Luo*[2],

[1]School of Languages and Linguistics, Griffith University, Brisbane, Australia
[2] Shenzhen Institute of Information Technology, China,

[1]c.tsurutani@griffith.edu.au, [2]luoda@sziit.com.cn

Abstract

This study aims to investigate native speakers' perception of prosodic variation of Chinese utterances. It is known that timing is crucial for intelligibility in English, Japanese and other accent languages [9, 13 ,14 ,16]. A tone language, Chinese relies heavily on the use of pitch more than do other languages for the purpose of distinguishing the meaning of the segmentally same words as well as expressing intonation. It is expected that pitch is the most important prosodic factor for the naturalness judgment of Chinese. However, no empirical data have been presented to support this view to date. In general, listeners are more sensitive to the deviation of timing than pitch. Pitch change also triggers slight change in duration. Whether the significance of timing for naturalness is universal across languages and applies to tone languages as well, relative importance of timing and pitch in Chinese was investigated using L2 Chinese speech. The results indicated that Chinese native listeners do notice the deviation of timing and regard it as accented speech.

Index Terms: naturalness, pitch, timing, foreign accent, Chinese

1. Introduction

The impact of prosodic features on naturalness has been acknowledged both in teacher belief and pronunciation research. Among world languages, three major prosodic features – timing, pitch and intensity – are coordinated to constitute the rhythm of languages by their phonological rules. However, the relative importance of each prosodic feature in L2 pronunciation has not been fully explored across languages. In L2 acquisition, prosody typically becomes the last hurdle to overcome. A study of L2 English reports the importance of temporal properties of speech [9,11]. Prosodic features are a key component of natural and intelligible speech, and thus need to be put under examination to find out exactly which features strongly affect native speakers' judgment of L2 speech. Many works were carried out in accent languages, such as English, which is a stress-accent language, and Japanese, which is a pitch accent language [9,14,15,16], while non-accent languages, namely tone-languages have not been studied well in this research paradigm.

In Chinese, tone is viewed as phonemic distinctions attached to the syllable. Tones are described by their F0 contour, amplitude and duration. While pitch plays an important role to indicate the patterns of tones and amplitude information was also found to be an useful cue to identify tones [6], timing is considered less important. The neglect of the role of timing is probably due to the syllable structure of Chinese which determines the vowel length. Vowels are long in open syllables and short in closed syllables. As short vowels in open syllables do not have tones, vowel length is predicted by whether the syllable carries tones or not.

Table 1: *Syllable structures of Chinese*

Syllable structure	Mora	Vowel length	Tone
(C)(G)VC	mm	short	Yes
(C)(G)VV	mm	long	Yes
(C)(G)V	m	short	No tone (neutral)

(G) = glide

Neutral toned syllables are not contained in contents words, which means that the prosodic words in Mandarin are all bi-moraic. The reason that duration is regarded as less important in Mandarin seems to be found in this morphological composition of Chinese words.

A bi-moraic foot, which consists of a heavy syllable or two light syllables, is the most common and stable prosodic structure found in world languages. Thus, it is not hard for learners of Chinese to maintain the appropriate vowel length for syllables.

However, whether native listeners of Chinese do consider the correctness of timing in the judgement of naturalness of Chinese is worthy of investigation to better understand the nature of foreign accents from phonetic and phonological perspectives and to address basic questions of speech science and technology in the domain of accent accommodation in spoken language comprehension.

In this study, the influence of prosodic features, timing and pitch in particular, will be examined, using native listeners' judgment of L2 Chinese utterances. L2 speech was extracted from German-speaking learners' utterances [5]. Like English, German is a stress-based, stress accent language. We should be able to obtain a sufficient amount of incorrect production from the speech data of learners whose L1 has a prosodic type different from the target language. The following sections discuss the role of timing in Chinese, the experiment and the result.

2. Prosodic features of Mandarin Chinese

Unlike Japanese or English, vowel length is not contrastive in Chinese. Due to the strong focus on tone, timing has been paid little attention in learning Chinese. It is true that pitch is a major prosodic component of tones. However, there is some evidence that implies the relevance of other prosodic factors to naturalness of Chinese.

The majority of Mandarin Chinese words (71%) are disyllabic and occasionally trisyllabic [1]. Mandarin Chinese has stress and a stressed syllable is produced with wider pitch range (pitch range expansion – [10]), longer duration, spectrum tilting and loudness, correlation of which is rather complicated

[8]. Among 20,000 disyllabic words and compounds, the following stress patterns for Chinese disyllabic words have been reported;

Table 2: *Stress pattern of disyllabic words*

Pattern	mora	Count	%
1.Heavy-Light (paa.pa) *dad*	M*m-m	1,500	7.5
2.Heavy-Heavy (tɕii.x^{w}aa) *plan*	Mm-mm	4,500	22.5
3. Heavy-(Heavy Ø) (s^{w}uu, s ɤ ɤ) *dorm*	mm-Mm	14,000	70.0
All		20,000	100.0

(Xu, 1982) [17]

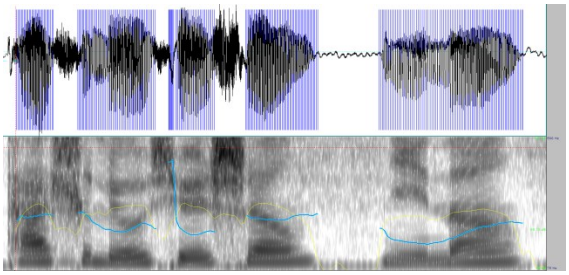
*M = stressed

The third pattern is the most common pattern in standard Mandarin. The first and second patterns represent trochees where the second syllable has a neutral tone in the first pattern and a full syllable in the second pattern while the third pattern shows iambs with two full-toned syllables. Native speakers find stress and stress levels notoriously difficult to judge. However, four stress levels are defined by the OSU system [10] as part of the attempt to create annotation conventions of Mandarin.

Table 3: *Stress levels in the stress tier* [10]

S3 syllable with fully realized lexical tone
S2 syllable with substantial tone reduction (e.g. undershooting of tonal target with duration reduction)
S1 syllable that has lost its lexical tonal specification (e.g. in a weakly stressed position)
S0 syllable with lexical neutral tone (i.e. such a syllable is inherently unstressed)

As shown in the table, different stress levels affect the duration of syllables. It is quite possible that native listeners recognize the deviation of timing at a particular level of stress.
The following spectrogram shows the duration of each syllable in Sentence ID 12.



ta shi liu xue sheng ni na
S3 S2 S3 S2 S3 S3 S0
156-118-203-174-325-149-274 (ms)

Figure 1: *Duration of syllables in the sentence '他是留学生, 你呢?'*

Among a seemingly simple succession of heavy syllables, the sentence exhibits vowel reduction after a stressed syllable and vowel lengthening at phrase and sentence final positions.
Mandarin Chinese is classified in the syllable timed language group according to the index based on the acoustic measurement of durations of vowels and intervocalic intervals [7]. However, it has been found that all languages in the world cannot fall into the rhythmic categories, - syllable-timed vs. stress-timed - as proposed by Abercrombie, and rigid isochrony in speech is lacking in both groups [2, 7]. Compared with Cantonese, Mandarin seems to present stress alternation [9], that is, stress-timed tendency as shown in the sentence above. Speech rhythm that we perceive could result from the combination of phonological, phonetic, lexical and syntactic facts. The perception native speakers have of their language, and the information on when they feel the deviation of rhythm, will provide a useful insight to the mechanism of Chinese prosody.
It can be predicted from the results of the previous studies that timing is the most important factor for assessment of L2 speech in many languages in the world. Tsurutani [14] examined Japanese native listeners' perceptions of L2 speech, using English-speaking learners' natural speech and found that correctness in timing was more important than that in pitch for the L2 speech to be perceived as more natural by Japanese native listeners.
In this study the same research design as the previous study [14] was employed to test the influence of timing and pitch on Chinese native listeners' judgment of L2 learners' utterances.

3. Experiment

3.1. Materials

The stimuli were selected from the database of German learners of Mandarin (DGM) collected at Free University of Berlin [5].
Sentences of 6-9 syllable length spoken by German students who had completed 12-28 weeks of Chinese language training at university were used, with consideration of the variety of tone change, difficulty of tone (tone sandhi, tone change in 'bu') and phonemes (j, q, x, sh, ch, zh, z, c, s) for L2 learners.
Utterances that contained timing errors or pitch errors or both, with no obvious segmental errors, were chosen, together with a correct utterance from a large sample of speech data.

The judgment of errors was made by three phonetically trained Chinese instructors and acoustic analysis.
The following four patterns were considered.

1. incorrect pitch, correct timing -- PxTo
2. correct pitch, incorrect timing -- PoTx
3. correct pitch, correct timing -- PoTo
4. incorrect pitch, incorrect timing -- PxTx

Using the same sentence would be convenient for analysis but could wear listeners' concentration. Thus, five sentences that contained the listed error patterns were chosen and formed 20 stimuli. The four patterns, 1- 4, of the same sentence were played, but the order of error patterns was randomized. Since the stimuli were natural utterances, it was not possible to control the degree of incorrectness as accurately as in synthesized speech. Efforts were made to select speech with similar speech rate and number of errors. Errors at both word

and phrase levels were counted. The following is the list of the sentences. The number of syllables and boundary tones are indicated at the end of the sentences

Table 4: *Stimulus sentences*

12 ta1 shi4 liu2 xue2 sheng1 ni3 ne0 ↗ 7
(S/he is an exchange student, How about you?)
他是留学生，你呢？

15 tian1 qi4 leng3 bu4 leng3 ↘ 5
(The weather is cold or not.)
天气冷不冷？

16 ni2 you3 mei2 you3 nan2 peng2 you0 ↘ 7
(Do you have a boy friend?)
你有没有男朋友？

22 ni2 zen3 me0 bu4 zhi1 dao4 a0 ↘ 7
(Why don't you know it?)
你怎么不知道啊？

14 wo3 shi4 yu3 yan2 xue2 yuan4 de0 xue2 sheng1 ↗ 9
(I am a student studying Linguistics subjects.)
我是语言学院的学生。

3-2.Participants

52 native listeners of Chinese (M= 26, F= 26) whose age ranged from 18 to 57 recruited in Brisbane (Australia) participated in the perception test in return for a small payment. In a previous study [13] it was found that native listeners' judgment for non-native speech did not differ between natives who live in their home country and those who reside overseas. In this study native listeners who were in the residential country of the researcher were employed as participants.

3-3.Procedure

Chinese native listeners were asked to listen to the stimuli and judge the naturalness of utterances using a Likert scale. The task was conducted on-line using the following instruction:

You will be listening to 20 short utterances spoken by non-native speakers. Judge the naturalness of the utterance by choosing the appropriate number from 1-7. Make sure to give a higher score---7 (native like) to a Good utterance and a lower score to a Bad utterance. There is no right and wrong answer. Don't listen to the same speech sample more than twice. Just follow your intuition as a native speaker.
Before the task, three practice sentences will be played

It took around 15 minutes for the participants to complete the task including the practice session with four sentences. The order of 24 sentences was automatically randomised in order to avoid order of presentation effect.

4. Results

The average and standard deviation of score obtained from the judgement by the Likert scale were calculated for 20stimuli. First of all, four error patterns were compared using the average score for 260 sentences (5 sentences for 52 participants) in order to see if the four error patterns were distinguished by the listener participants. Standard deviations are listed below in Figure 1.

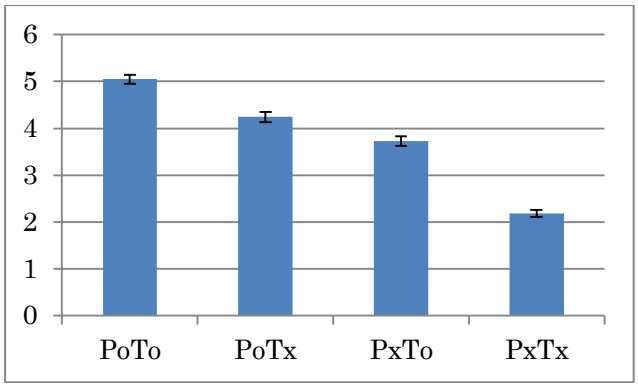


Figure 2: *The average score (+/- standard error) for four types of errors*

The average scores received were highest in sentences with PoTo, followed by PoTx and PxTo, and lowest in sentences with PxTx-type errors. This means that participants, on average, judged utterances with pitch errors as worse than those with timing errors. From the result shown in Figure 1, however, it is apparent that judgement of error types O and X will be different, but there appears to be likely overlap in the ranking of pitch and timing errors. The relative ranking of errors by individual participants was then looked at to further understand whether individuals more specifically distinguished between all four error types. Paired T-tests were undertaken to determine in particular if there were significant differences in rankings between the four error types. (Table 1)

Table 5**:** *Results of multiple t-tests (2 tailed tests) between four error types*

Comparisons	*/t/*	*p*	*t Critical*
PoTo-PoTx	*3.098*	*.002*	*1.963835*
PoTx-PxTo	*4.465*	*.000*	*1.963841*
PxTo-PoTo	*7.642*	*.000*	*1.963854*
PxTx-PoTo	*18.502*	*.000*	*1.964091*
PxTx-PoTx	*15.051*	*.000*	*1.964077*
PxTx-PxTo	*10.862*	*.000*	*1.963932*

This analysis showed significant differences not only between the most extreme assessment, but also between all other combinations, supporting the initial view of the error hierarchy presented in Figure 2.
In order to see the influence of error types, the stimuli were listed in the descending order of their average score by 52 participants together with the number of errors (see Table 6).
Stimuli at the bottom of the table had more pitch errors. It means that Native Chinese listeners are very sensitive to tone errors, but timing errors are sometimes tolerated. As the number of speech samples was limited, differences between error types were small in some sentences. To deal with this issue, a larger scale of data collection is currently under way.

5. Conclusion

The significance of timing in naturalness judgements was tested in Mandarin Chinese. When native listeners hear their

language spoken, they judge the naturalness of the rhythm as a whole. However, when we decompose the prosodic features, we can identify the relative importance of each component.

Table 6: *Average scores of each stimuli*

Sentence ID No.	Error pattern	Number of errors		Score
		Timing	Pitch	
14	PoTo			6.55
16	PoTx	1		6.17
14	PxTo		2	5.41
22	PoTo			5.25
16	PoTo			5.15
22	PoTx	1		4.77
12	PoTx	2		4.31
15	PoTo			4.18
12	PoTo			4.13
15	PxTo		1	4.06
14	PoTx	2		3.48
22	PxTo		3	3.37
12	PxTo		1	3.17
16	PxTo		4	2.63
16	PxTx	2	3	2.62
15	PxTx	1	3	2.60
15	PoTx	1		2.48
22	PxTx	3	4	2.19
14	PxTx	1	3	1.82
12	PxTx	1	3	1.67

Among prosodic features, it was found that pitch was a more crucial factor than timing for naturalness judgement by Chinese native listeners. This result supports the general view that pitch is the most important prosodic feature in a tone language. However, there were significant differences in scores between all four error patterns. This means that the correctness of timing does play a crucial role as much as the correctness in pitch. In general, people can easily and accurately tell whether a sound in their native tongue was short or long, but cannot as easily determine pitch. In the case of Mandarin Chinese, native listeners have a sensitive and critical ear for the deviation of pitch due to the nature of tone languages. However, timing did matter in the judgement of naturalness as shown in the significant differences between PoTo - PoTx and PxTo - PxTx.

In conclusion, the significance of timing is universal and consequently found in also in Mandarin Chinese, a language which does not have length contrast between short and long vowels. Both types of prosodic errors, pitch and timing were detected by native listeners' critical listening (?) even in a tone language although the degree of significance of two types of errors differed from other languages. When Mandarin Chinese is taught to second language learners, attention will need to be given to timing as well as to pitch. Mandarin Chinese has stress and the duration of syllable varies depending on the stress level. The result of this study could be due to the fact that Mandarin Chinese is categorized as a stress-timed language. To claim the result of this study is applicable to the Chinese language, Cantonese, which does not have stress needs to be tested as well. It would be also interesting to expand the scope of research to other languages which do not have a clear rhythmic category, such as Korean.

6. Acknowledgements

We would like to thank Hansjoerg Mixdorff and his team who kindly provided their speech data for this study. This study was supported jointly by Griffith University Research Grant 2012, the Social Science Foundation of China (10CYY009), and the key project from Jiangsu Higher Education Institutions' Key Research Base for Philosophy and Social Sciences (2010JDXM024).

7. References

[1] Dyanmu, S. (2007) *The phonology of standard Chinese,* Oxford University Press

[2] Dauer, R.M. (1983) Stress-timing and syllable-timing reanalyzed, *Journal of phonetics,* 11 56-62.

[3] Beckman, M. and Pierrehumbert, J. (1986). "Intonational structure in Japanese and English", *Phonology Yearbook.* 155-300.

[4] Cheng, J. (2011). Automatic Assessment of prosody in high-stakes English tests *INTERSPEECH* 2011, 1589-1592

[5] Chiu, C-Y, Liao, Y-F, Mixdorff, H., Chen, S-L (2009) A preliminary study on corpus design for computer-assisted German and Mandarin language learning Speech Database and Assessments, 2009 Oriental COCOSDA International Conference, 154 – 159.

[6] Coster, D.C., & Kratochvil, P. (1984) Tone and stress discrimination in normal Peking dialect speech. In B.Hong (ed.) New papers in Chinese linguistics 119-132, Canberra; Austalian National University Press

[7] Grabe,E and Low, L (2002) Durational variability in speech and the ehythm class hypothesis in Laboratory Phonology 7, 515-546

[8] Li, A. (2002). Chinese prosody and prosodic labeling of spontaneous speech, Speech prosody 2002, France

[9] Maier, A., Honig,F.,Zeissler, V,, Batliner, A. ,Korner,E., Yamanaka, N.,Ackermaqnn, P. and North, E. (2009). A Language-independent feature set for the automatic evaluation of prosody. *Interspeech 2009*, 600-603.

[10] Shu-hui Peng, Marjorie K.-M. Chan, Chiu-yu Tseng, Tsan Huang, Ok JooLee, and Mary E. Beckman (2005). Towards a Pan-Mandarin system for prosodic transcription. In Sun-Ah Jun, ed. Prosodic Typology: The Phonology of Intonation and Phrasing, pp. 230-270. Oxford, UK: Oxford University Press.

[11] Tajima, K. Port, R. and Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English *Journal of Phonetics*, 25, 1-24.

[12] Tsurutani, C. (2008). *Pronunciation and Rhythm of Japanese as a second language* Hiroshima, Keisuisha.

[13] Tsurutani, C. (2009) Intonation of Japanese sentences spoken by English speakers *INTERSPEECH 2009*, 692-695.

[14] Tsurutani, C. (2010). Foreign accent matters most when timing is wrong, *Interspeech 2010* 1854-57.

[15] Tsurutani, C. (2011) "L2 Intonation - A study of L1 transfer in Japanese intonation by English-speaking learners", *Second Language,* 79-102.

[16] Warren, P. Elgort,I. & Crabbe, D. (2009). Comprehensibility and prosody ratings for pronunciation software development, *Language learning & Technology* Vol. 13, No.3, 87-102.

[17] Xu, S. (1982) 'Shuangyinije ci de yinliang fenxi' (A quantitative analysis of disyllabic words) *Yuyan Jiaoxue Yu Yanjiu* 2, 4-19