Vowel identity conditions the time course of tone recognition

Jason A. Shaw[1,3], *Michael D. Tyler*[1,2], *Benjawan Kasisopa*[1], *Yuan Ma*[3], *Michael Proctor*[1,3], *Chong Han*[3], *Donald Derrick*[1], *Denis Burnham*[1]

[1] MARCS Institute, University of Western Sydney, Australia
[2] School of Social Sciences and Psychology, University of Western Sydney, Australia
[3] School of Humanities and Communication Arts, University of Western Sydney, Australia

J.Shaw@uws.edu.au, M.Tyler@uws.edu.au, B.Kasisopa@uws.edu.au, 16603080@uws.edu.au, Michael.Proctor@uws.edu.au, C.Han@uws.edu.au, D.Derrick@uws.edu.au, Denis.Burnham@uws.edu.au

Abstract

Using eye-tracking in a visual world paradigm, we sought converging evidence for the time course of Mandarin Chinese tone recognition as predicted by the availability of information in f0 and past results from a gating experiment. Our results showed that tones 1 and 2 are recognized earlier than tone 4, followed by tone 3. With the exception of tone 2, which was recognized earlier than expected, our results are consistent with those found in gating. The speed of tone 2 recognition varied significantly across vowels in our study, part of a broader pattern whereby vowels systematically influenced the time course of tone recognition. Rising tones, tone 2 and tone 3, were recognized earliest when co-produced with /a/. The falling tone, tone 4, was recognized earliest when co-produced with /u/. Intrinsic f0 and spectral cues to tone are discussed as possible explanations for the vowel quality effect.

Index Terms: speech perception, word recognition, tone, visual world, eye-tracking, Chinese, intrinsic f0

1. Introduction

The time scale over which phonological information becomes available in speech varies across phonological categories. The lexical tones of standard Chinese (Mandarin) offer clear examples of categories that distribute information across a syllable. Acoustic parameters that differentiate Mandarin tones are well-known [e.g., 1, 2, 3] and include those available very early in the temporal unfolding of the syllable (e.g., f0 onset), those computed over the entire syllable (e.g., average f0, syllable duration, amplitude contour), and those computed over some sub-interval of the syllable (e.g., f0 slope, f0 turning point). As such, lexical tone offers a window on how information distributed across time is integrated in perception.

Information about lexical tone available in f0 unfolds at different rates across the four lexical tones of standard Chinese. On the basis of these differences, Lai and Zhang [4] predict that the time course of word recognition should differ across tones. Tone 1 of Mandarin is a high, level tone; Tone 2 starts mid and rises; Tone 3 starts low and rises; Tone 4 starts high and falls. Following the temporal availability of information in f0, Lai and Zhang predict that tone 1should be recognized earliest followed by tone 4, tone 2, and then tone 3. They provided evidence for this order of recognition using a gating experiment.

In a gating experiment, participants are asked first to identify a word or tone from a short auditory stimulus, a recording of the target "trimmed" to include only the onset of the word. In subsequent trials, participants are played progressively longer chunks, or "gates", until the word can be accurately identified. One issue with the use of gating paradigms, however, is that more temporal information does not always lead to improved identification performance. For example, a gating paradigm with French listeners played words beginning with /tl/ and /dl/ sequences, which are not found in French [5]. On early gates, participants correctly identified the initial consonant as /t/ and /d/; however, as the number of gates increased, listener percepts switched to /kl/ and /gl/, sequences that are phonotactically licit in French. This indicates that longer chunks of speech may engage different types of perceptual processes. Moreover, perceptual processes operating over longer time windows may override or otherwise influence early information. Specifically with respect to tone, some perceptual models posit that information is processed differently, depending upon where it occurs in the syllable [6]. The abbreviated stimuli used in gating may potentially obscure natural integration of temporally distributed information.

In this study, we sought converging evidence for the time course of tone recognition in an eye-tracking study using a variant of the visual world paradigm [e.g., 7, 8, 9]. This paradigm allows assessment of the time course of tone recognition with naturalistic auditory stimuli. In addition to seeking converging evidence from a new paradigm, we also added a new manipulation. We evaluated whether vowel quality influences the time course of tone perception. The influence of vowel quality on f0 has been argued to be universal [10]. Across dialects of Chinese, vowels and tones interact at both sub-phonemic [11] and categorical [12] levels of description. Specifically in Mandarin Chinese, production studies have shown adjustments of vowel articulation according to tone [13, 14]. However, it is not known whether observed variation in production has a significant impact on the time course of tone recognition. To investigate this, our stimulus materials crossed the four Mandarin tones with three different vowels.

2. Experimental Methods

2.1. Participants

The sample of listeners consisted of 16 native speakers of Mandarin Chinese (2 male) ranging from 19 to 47 (mean = 24; stdev = 6) years of age recruited from the UWS community. Two additional participants were tested but their data were discarded due to technical problems with data acquisition.

2.2. Procedure

Participants were seated in front of a computer monitor in a quiet room. They positioned their chin and forehead on a chin

rest located 50cm from the monitor. Auditory stimuli were played over loudspeakers located next to the monitor. A Tobii-x120 eye tracker sampled eye-movements at 60 Hz. Before beginning the practice section, the eye-tracker was calibrated to the gaze of each participant.

Figure 2 shows the trial procedure. Participants were first shown a preview of four words on the screen (Figure 2, left panel). They were instructed to read the words, click on the crosshair in the center of the screen, and then click on the word that they heard over the loudspeaker. After clicking on the crosshair, a red triangle appeared around the cross-hair (Figure 2, middle panel). When continuous gaze duration to the crosshair reached 200ms, the crosshair and red rectangle disappeared and the auditory stimulus played (Figure 2, right panel). This procedure allowed participants to preview words in the corners of the screen and ensured that they were fixating on the center of the screen at the start of the auditory stimulus.

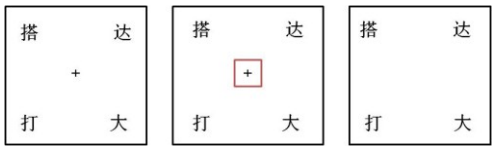


Figure 2: *Schematic depiction of an experimental trial (test condition). Participants clicked on the crosshair, left panel; fixated on the cross hair, middle panel; and clicked on the word that they heard, final panel.*

2.3. Materials

2.3.1. Visual stimuli

The visual stimuli consisted of 28 characters representing 28 distinct syllables of Mandarin. The syllables were made up of the vowels /a/, /i/ and /u/ combined with all four tones and /b/, /d/, and /t/ onsets. A complete list of stimulus items is given in Table 1.

Table 1: *Stimulus items—Romanization is provided below each Chinese character.*

	/a/	/i/	/u/
Tone 1	搭(da)八(ba)	低(di) 逼(bi) 梯(ti)	都(du)突(tu)
Tone 2	达(da)拔(ba)	敌(di)鼻(bi)提(ti)	读(du)图(tu)
Tone 3	打(da)把(ba)	底(di)比(bi)体(ti)	赌(du)土(tu)
Tone 4	大(da)爸(ba)	地(di)必(bi)替(ti)	度(du)兔(tu)

2.3.2. Auditory stimuli

The auditory stimuli consisted of all 28 syllables shown in Table 1. These syllables were recorded by three female native speakers of Mandarin Chinese from northern China. Two tokens of each syllable per speaker were included in the experiment yielding a total of 168 sound files. The individual sound files were excised from 100 ms before the onset of speech until the end of the utterance and a 20 ms onset and offset ramp was applied. Tokens were normalized to 90% of the peak amplitude (across all tokens) and a 10Hz 6th-order ellipitical high-pass filter was applied to remove the DC component. Each sound file was played four times for a total of 672 trials (672 = 28 syllables * 3 speakers * 2 tokens * 4 repetitions), which were played at random. Participants were given a break half-way through the experiment.

2.4. Conditions

The auditory and visual stimuli were organized into two conditions, a test condition and a control condition. In the test condition, the target word was shown on the screen (in one of four corners) along with competitor words differing only in tone (see Figure 3 for an example). In the control condition, the target word appeared on the screen with competitors that differed on the onset consonant as well as on the vowel and/or tone. The purpose of the control condition was to provide baseline word-recognition data for each token in the experiment so that differences in target fixation patterns across conditions could be attributed to the role of tone in word recognition.

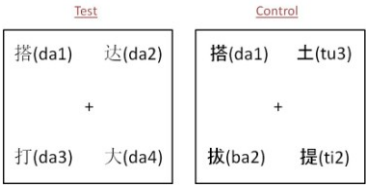


Figure 3: *Examples of test (left) and control (right) conditions. In these examples, 搭(da1), is the target. In the test condition, competitors differ only in tone. In the control condition, competitors differ in both the onset consonant AND either vowel or tone.*

Trials containing each sound file occurred twice in the test condition and twice in the control condition. The distribution of auditory and visual stimuli was balanced so that each character appeared as target and as distracter equal numbers of times, each character occurred in test and control conditions equal numbers of times, and each character occurred in each corner of the screen equal numbers of times. The presentation of all stimuli, including control and test conditions, was fully randomized for each participant.

2.5. Data preparation and analysis

The dependent variable in the analysis was participant looks to the target word, or target fixation. Since the same auditory stimuli were presented in both control and test conditions, differences in target fixation across conditions can only be attributable to the competitor population (tone competitors in the test condition vs. dissimilar competitors in the control condition). To compute target fixation, we coded eye fixations for area of interest, either a fixation to target, 1, or a fixation elsewhere, 0, for each sample of data (16.67 ms intervals).

To correct for eye-movement-based dependencies, the binary data was aggregated into 50 ms bins (three samples per interval). The empirical logit (elogit) and associated weights were calculated over trials within bin, condition, subject, and item. The elogit transformation converts fixation proportions to a continuous scale without upper/lower bounds. Since our design includes multiple repetitions of stimulus items, we were able to compute the elogit within subjects and items and retain the possibility of computing crossed random effects in a linear regression model (see results).

We determined the time window for analysis by plotting the elogit across all conditions as a function of time and identifying (1) a sharp increase in looks toward target and (2) a plateau in looks towards target. This was done across conditions to eliminate hypothesis-based bias in window selection [15]. The selected analysis window begins at 300 ms

and ends at 800 ms. The onset of 300 ms is reasonable, since there was 100 ms of silence before each sound file and 200 ms is roughly the time required to plan an eye-movement [16]. The duration of the window is greater than the duration of the stimuli, which ranged between 250-500ms, and therefore can reflect looks driven by phonologically relevant information distributed across the word.

3. Results

3.1. Visualization

The top panel of Figure 4 shows target fixation (elogit transformed) across conditions (separate lines) for each tone. For all tones, there were fewer target fixations in the test condition than in the control condition indicating that, as expected, the time course of syllable recognition is slowed when there are lexical tone competitors. The degree to which tone impacts word recognition is represented in the differences between the lines (control vs. test) for each tone. These differences tend to be concentrated in the early part of the syllable. Figure 5 collapses across time to show the mean difference (between target and control conditions) in target fixation for each tone. The difference appears to be smallest for tone 2 followed by tone 1, tone 4 and tone 3.

Figures 6 and 7 break the data down by tone-vowel combination. It is clear, firstly, that the patterns in Figure 4 are not uniform across vowels. For example, early recognition of tone 2 (Figure 5) is driven by /a/, which shows the smallest difference between conditions (Figure 7). Early in the time window, tone 3, the other rising tone, is also recognized quickly when co-produced with /a/ (Figure 6). Tone 4, on the other hand, is recognized earliest when co-produced with /u/.

Figure 4: *Target fixation (y-axis) by time (x-axis) for each tone (rows). Lines show test and control trials.*

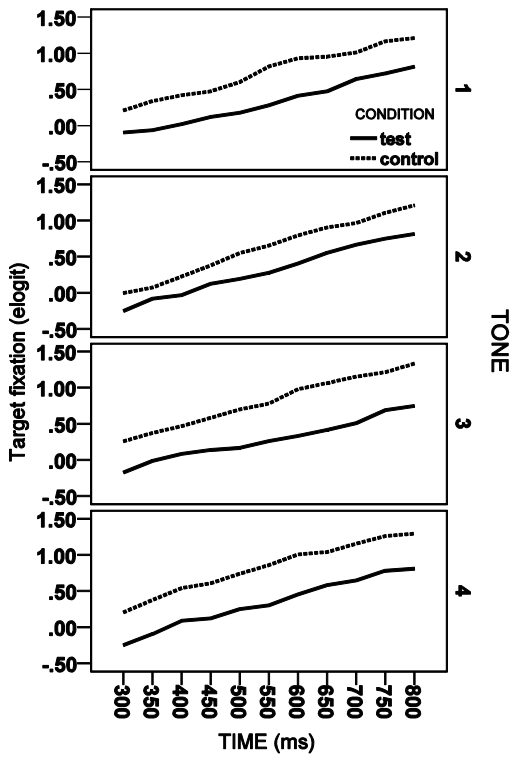


Figure 5: The *difference in target fixation (y-axis) between test and control conditions averaged across 300-800 ms for each tone (x-axis).*

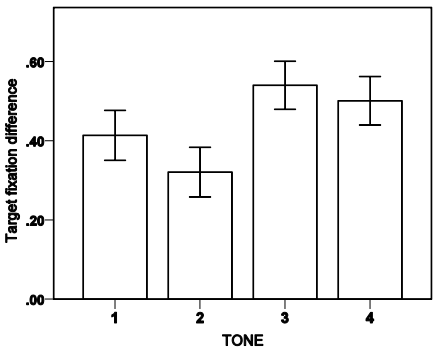


Figure 6: *Target fixation (y-axis) by time (x-axis) for each tone (rows) and vowel (columns) combination. Lines show test and control trials.*

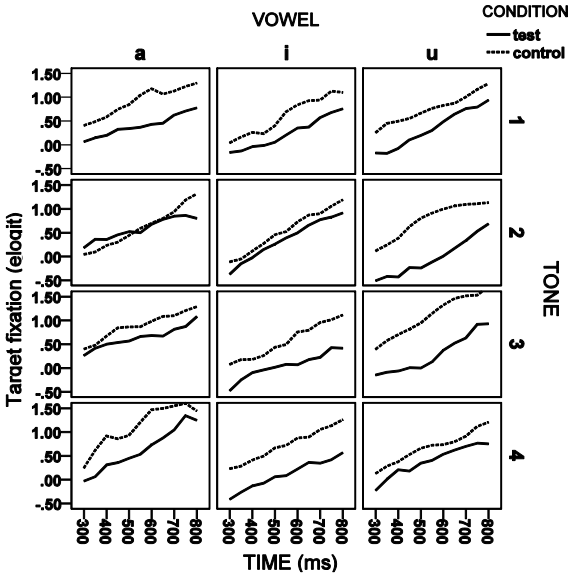
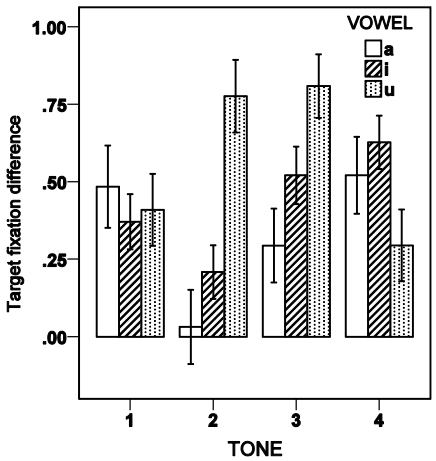


Figure 7: *difference in target fixation (y-axis) between test and control conditions across 300-800 ms by tone (x-axis) and vowel (separate bars) combination.*



3.2. Model

To account for fixation patterns, we fit a mixed effects linear regression model to the elogit-transformed data, with crossed random effects for subject and item. Vowel {a, i, u}, tone {1, 2, 3, 4}, condition {control, test}, time {50ms bins}, and all interactions between them were coded as fixed effects. The log character frequency of target items was also included as a fixed effect in an initial model but was eliminated from the final model because its inclusion did not significantly improve the model's fit and it did not interact with other factors. The model was coded in R [17] using the *lme4* package [18]. Table 2 summarizes model parameters. Reported *p*-values were obtained by treating the *t*-values as if they were drawn from a normal distribution, following a suggestion by Barr (http://talklab.psy.gla.ac.uk/tvw/elogit-wt.html). The inclusion of time as a fixed factor is an important innovation in the multi-level logistic regression framework for analysis of the visual world paradigm [15]. By including time as a fixed factor, we evaluate explicitly whether our experimental manipulations affected the time course of processing.

Table 2. *Fixed factors and interactions in mixed effects model of target fixation (elogit).*

Effect	β	SE(β)	*t*	*p*
Intercept	-0.13	0.31	-0.4	0.68
Condition(Cond)	-1.35	0.24	-5.7	0.00
Time	0.74	0.43	1.7	0.09
Tone	-0.05	0.11	-0.4	0.66
Vowel	-0.24	0.13	-1.8	0.08
Cond*Time	1.94	0.44	4.4	0.00
Cond*Tone	0.46	0.09	5.2	0.00
Time*Tone	0.29	0.15	2.0	0.05
Cond*Vowel	0.66	0.11	6.0	0.00
Time*Vowel	0.49	0.19	2.6	0.01
Tone*Vowel	0.04	0.05	0.8	0.44
Cond*Time*Tone	-0.57	0.16	-3.5	0.00
Cond*Time*Vowel	-0.74	0.20	-3.6	0.00
Cond*Tone*Vowel	-0.19	0.04	-4.6	0.00
Time*Tone*Vowel	-0.15	0.07	-2.1	0.03
Cond*Time*Tone*Vowel	0.25	0.07	3.4	0.00

The main effects of vowel ($p = .08$) and tone ($p = .66$) on target fixation were not significant, nor was the interaction between vowel and tone ($p = .44$). This was expected. As our design involves a comparison between test and control conditions, we are interested in factors that modulate the effect of condition. Condition was significant ($p < .001$) as was the vowel*condition interaction ($p < .001$) and tone*condition interaction ($p < .001$). The three-way interaction between condition, tone and vowel was also significant ($p < .001$). These effects reflect differences shown in Figure 5 and Figure 7. They can be viewed as "whole window" effects because they are computed across the entire analysis window. However, each of these effects, vowel*condition, tone*condition, vowel*tone*condition, interacted with time in our regression model. This indicates, as shown in Figures 4 and 6, that the whole window effects are concentrated in sub-regions of the analysis window.

To explore the interaction between tone and condition, we fit mixed-effects regression models to each pairwise combination of tone. The interaction between tone and condition was not significant for the comparison of tone 1 and tone 2. All other comparisons were significantly different at the $p < .01$ criterion. Tone 1 and tone 2 showed the smallest modulations of condition (indicating early recognition) and modulated condition significantly less than tone 4. Tone 4 modulated the effect of condition less than tone 3. The resulting hierarchy of tone recognition is as follows: 1 = 2 > 4 > 3, where ">" indicates "is recognized earlier than".

4. Discussion

Using a visual world paradigm, we investigated the time course of tone perception in Mandarin Chinese. In light of past results from gating paradigms and f0 patterns known to differentiate tones, we expected that tone 1 would be recognized earlier than tone 4 followed by tone 2 and then tone 3 (tone 1 > tone 4 > tone 2 > tone 3). In our data, tones 1, 3, and 4 followed the expected pattern, but tone 2 was recognized earlier than expected.

We also found that tone recognition, particularly for dynamic tones, varied substantially across vowels. This variation could be responsible, at least in part, for the earlier than expected recognition of tone 2. Tone 2 recognition was particularly early for /a2/. This was part of a broader generalization in our data—/a/ facilitated early recognition of rising tones; /u/ facilitated early recognition of falling tones. This pattern is possibly related to the intrinsic f0 of these vowels. Intrinsic f0 is lowest for /a/ and highest for /u/ [10]. The slightly lower and higher f0 starting points for /a/ and /u/, respectively, stretch the range of f0 early in the syllable. This may facilitate early recognition of dynamic tones. For falling tones, the higher f0 of /u/ may highlight the characteristic decrease in f0. Likewise for rising tones, the lower f0 of /a/ may highlight the characteristic rise in f0.

Another possible explanation for the effect of vowel quality on the time course of tone recognition is that information about lexical tone is present in other acoustic dimensions, besides f0. Erickson *et al.* [14] found that when tones 1 and 3 were co-produced with /a/, tongue height was lower (and F1 higher) for tone 3 than for tone 1. These kinds of effects are not necessarily reliable across speakers and vowels [19], but listeners may make use of them when they are available or when f0 information is not available [20].

Overall, our results show that the visual world paradigm can offer insight into the time course of tone processing using natural auditory stimuli. With the exception of tone 2, our results replicated those obtained in Lai and Zhang's [4] gating task. The different results for tone 2 across experiments are likely due, at least in part, to the particular vowels with which tone 2 was co-produced. In our study, tone 2 was recognized very quickly when co-produced with /a/ and slowly when co-produced with /u/. Conclusions about the time course of tone processing are therefore most conservative when restricted to tone-vowel combinations. Additional research is necessary to determine the degree to which intrinsic f0 and/or tonal information distributed across other acoustic dimensions influences the time course of tone recognition.

5. Acknowledgements

This research was supported by an internal grant from the MARCS Institute to authors Han, Derrick, Proctor, Shaw, Tyler. We'd like to thank Jia Ying for help recruiting and administering the experiment and San Duanmu, Wei-rong Chen, Allard Jongman and Joan Sereno for helpful comments and discussion.

6. References

[1] J. Gandour, "Tone perception in Far Eastern languages," J Phonetics, vol. 11, pp. 149-175, 1983.

[2] C. B. Moore and A. Jongman, "Speaker normalization in the perception of Mandarin Chinese tones," J. Acoust. Soc. Amer., vol. 102, pp. 1864-1877, Sep 1997.

[3] D. H. Whalen and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica, vol. 49, pp. 25-47, 1992.

[4] Y. Lai and J. Zhang, "Mandarin lexical tone recognition: The gating paradigm," Kansas Working Papers in Linguistics, vol. 30, pp. 183-194, 2008.

[5] P. A. Hallé, et al., "Processing of illegal consonant clusters: A case of perceptual assimilation?," J. Exp. Psych.: Hum. Percept. & Perform., vol. 24, p. 592, 1998.

[6] D. House, "Differential perception of tonal contours through the syllable," in Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, 1996, pp. 2048-2051.

[7] J. M. Mcqueen and M. C. Viebahn, "Tracking recognition of spoken words by tracking looks to printed words," The Quarterly Journal of Experimental Psychology, vol. 60, pp. 661-671, 2007.

[8] M. K. Tanenhaus, et al., "Integration of visual and linguistic information in spoken language comprehension," Science, vol. 268, pp. 1632-1634, Jun 16 1995.

[9] A. Weber and A. Cutler, "Lexical competition in non-native spoken-word recognition," J. Mem. & Lang., vol. 50, pp. 1-25, Jan 2004.

[10] D. H. Whalen and A. G. Levitt, "The universality of intrinsic F0 of vowels," Journal of Phonetics, vol. 23, pp. 349-366, 1995.

[11] E. Zee, "Tone and vowel quality," Journal of Phonetics, vol. 8, pp. 247-258, 1980.

[12] I. Maddieson, "The Intrinsic Pitch of Vowels and Tones in Foochow," University of California Working Papers in Phonetics, vol. 33, pp. 191-202, 1976.

[13] P. Hoole and F. Hu, "Tone-vowel interaction in standard Chinese," in International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, 2004.

[14] D. Erickson, et al., "Effect of tone height on jaw and tongue articulation in Mandarin Chinese," in International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, 2004.

[15] D. J. Barr, "Analyzing 'visual world' eyetracking data using multilevel logistic regression," J. Mem. & Lang., vol. 59, pp. 457-474, Nov 2008.

[16] E. Matin, et al., "Saccadic overhead: Information-processing time with and without saccades," Attention, Percept., & Psychophys., vol. 53, pp. 372-380, 1993.

[17] R Core Team, "R: A language and environment for statistical computing, version 2.15.3" ed: R Foundation for Statistical Computing, 2013.

[18] D. Bates, et al., lme4: Linear mixed-effects models using S4 classes. R package version 0.99999-0, 2012.

[19] P.-C. Torng, "Supralaryngeal articulator movements and laryngeal control in Mandarin Chinese tonal production," University of Illinois at Urbana-Champaign, 2000.

[20] S. Liu and A. G. Samuel, "Perception of Mandarin lexical tones when F0 information is neutralized," Language and Speech, vol. 47, pp. 109-138, 2004.