JASA ARTICLE

Effects of vowel coproduction on the timecourse of tone recognition

Jason A. Shaw[1,a)] and Michael D. Tyler[2,b)]
[1]Department of Linguistics, Yale University, Dow Hall, New Haven, Connecticut 06511, USA
[2]School of Psychology, Western Sydney University, Penrith, New South Wales 2751, Australia

ABSTRACT:
Vowel contrasts tend to be perceived independently of pitch modulation, but it is not known whether pitch can be perceived independently of vowel quality. This issue was investigated in the context of a lexical tone language, Mandarin Chinese, using a printed word version of the visual world paradigm. Eye movements to four printed words were tracked while listeners heard target words that differed from competitors only in tone (test condition) or also in onset consonant and vowel (control condition). Results showed that the timecourse of tone recognition is influenced by vowel quality for high, low, and rising tones. For these tones, the time for the eyes to converge on the target word in the test condition (relative to control) depended on the vowel with which the tone was coarticulated with /a/ and /i/ supporting faster recognition of high, low, and rising tones than /u/. These patterns are consistent with the hypothesis that tone-conditioned variation in the articulation of /a/ and /i/ facilitates rapid recognition of tones. The one exception to this general pattern—no effect of vowel quality on falling tone perception—may be due to fortuitous amplification of the harmonics relevant for pitch perception in this context.
 https://doi.org/10.1121/10.0001103

(Received 10 June 2019; revised 29 March 2020; accepted 30 March 2020; published online 23 April 2020)
[Editor: Megha Sundara] Pages: 2511–2524

I. INTRODUCTION

A. Independence of tones and vowels in perception

In speech perception, vowel contrasts tend to be perceived independently of pitch (e.g., Poeppel *et al.*, 1997). The formants that differentiate vowels in acoustics are structured by articulatory gestures that condition changes in vocal tract shape over time. Resonances of the vocal tract characteristic of a particular vowel are, by definition, constant across variations in the source of acoustic excitation. In perceiving vowels across variation in pitch, listeners would seem to be picking out aspects of the signal related to vocal tract resonance. While listeners tend to experience perceptual constancy for vowels across variation in pitch, it is not known whether listeners also experience perceptual constancy for pitch across variation in vowel quality. We investigated this issue in the context of a lexical tone language, Mandarin Chinese.

In lexical tone languages, pitch is often perceived categorically in terms of tone (Peng *et al.*, 2010; Xu *et al.*, 2006), particularly for dynamic tones, i.e., those that change in pitch over time (cf. Abramson, 1977; Sun and Huang, 2012). Moreover, vowels and tones are functionally equivalent from a linguistic standpoint in that both are used to differentiate word meanings. For example, changes in tone from a rising tone /ti35/ to falling tone /ti51/ affect a change in meaning in the same way as a change in vowel from, e.g., /ti35/ to /ta35/. Accordingly, changes to vowels and tones resulting in semantic anomaly evoke comparable neural responses (Schirmer *et al.*, 2005). Cues to vowels and tones are simultaneously (or near simultaneously) present in the acoustic signal. This is because the acoustic energy that typically excites vowel resonances in the vocal tract has its source in vocal fold vibration, the fundamental frequency of which cues pitch perception. At issue is whether listeners perceive pitch independently of the vowels that shape vocal tract resonance.

To the extent that fundamental frequency, f_0, is lower than F_1, as is typical for adult speech, we may expect that tone perception is entirely unaffected by the spectral qualities of the vowel with which it is coproduced (and the associated supralaryngeal resonances, f_R).[1] While it is known that spectral qualities associated with voice quality influence pitch perception (e.g., Kuang and Liberman, 2018), listeners may be able to compensate in perception for the spectral qualities of vowels when they function as independent linguistic units. Unlike voice quality and pitch perception, which are known to covary (Kuang, 2017), there are few cases of tone-vowel covariation, and putative cases may be mediated by other factors such as syllable duration or voice quality. For example, some languages have "register" contrasts, which include aspects of pitch, voice quality, and vowel quality (e.g., Edmondson and Gregerson, 1993; Edmondson *et al.*, 2001); acoustic analyses have, in some cases, concluded that the observed differences in vowel quality are small and unlikely to be robust perceptual cues

a)Electronic mail: Jason.Shaw@yale.edu, ORCID: 0000-0002-5285-5049.
b)Also at: The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Penrith NSW 2751, Australia.

(Brunelle, 2009). Nevertheless, these are not cases in which contrastive tones and vowels overlap. Rather, in register languages, covariation in pitch, voice quality, and vowel quality cue a single contrast, possibly a constriction in the lower pharynx between the tongue root and the upper valves of the larynx (Edmondson *et al.*, 2001). Putative tone-vowel covariation may also arise from variation in duration. Shorter syllables may have the effect of reducing pitch range and also conditioning undershoot of vowel targets, leading to the appearance of covariation between tone and vowel quality (e.g., Jiang-King, 1999). Similarly, manipulation of vowel duration is known to influence the perception of tone (e.g., Greenberg and Zee, 1979), while a loss of vowel duration contrast can condition diachronic changes to tone systems (Gandour, 1977). These examples of tone-vowel interactions are indirect. As Yip (2002, p. 32) observed: "Direct and unequivocal interaction between vowel height and tone is extremely rare, and the examples are controversial. If a connection exists, one would expect higher vowels to be found with higher tones since in many languages, including English, studies have shown that higher vowels have an intrinsically higher fundamental frequency." In our discussion below, we take up the possible effect suggested by Yip (intrinsic f_0), as well as some other ways in which vowel quality could conceivably influence tone perception.

Phonetic descriptions of lexical tone are typically expressed as f_0 measurements over time and typically normalized by speaker (Abramson, 1962; Howie, 1976; Rose, 1988). In these descriptions higher harmonics are redundant. In pitch perception, however, higher harmonics are important. Minimally, models of pitch perception (e.g., Meddis and O'Mard, 1997) must account for results from the missing fundamental paradigm. Human listeners, as well as nonhuman primates (Tomlinson and Schwarz, 1988), cats (Heffner and Whitfield, 1976), and birds (Cynx and Shapiro, 1986), perceive pitch even when the fundamental frequency has been excised from the signal. This indicates that pitch perception involves integration of acoustic energy from multiple harmonics distributed across frequency bands. When perceiving speech, the structure of the harmonics that are relevant for pitch perception is filtered by the vocal tract in vowel-specific ways. This relationship raises the question of whether vowel identity might influence pitch perception.

How might the spectral filtering of vowel-specific vocal tract configurations impact pitch perception in speech? One possibility is that human listeners perceptually compensate for vowel quality, recovering harmonic structure from the acoustic signal despite the influence of vocal tract resonance. In this case, we would expect vowel quality to have negligible effects on tone perception. Possible evidence against this comes from an interference task (Garner paradigm), showing that vowel variation interferes with tone identification to a greater degree than tone variation interferes with vowel identification (Tong *et al.*, 2008). Another possibility is that listeners rely less on the second and higher harmonics in tone perception than in the perception of pitch more generally (i.e., non-speech). Kuang and Liberman (2018) showed that for English listeners spectral properties had a stronger influence on pitch perception for non-speech stimuli than for speech stimuli. Developmentally, one aspect of acquiring language-specific speech perception involves tuning in to the particular aspects of the signal that reliably cue phonological contrast. For lexical tone languages, this could mean enhanced perceptual attunement to f_0 to the exclusion of higher level harmonics since the amplitude of the harmonics will be perturbed by voice quality and vowel-specific resonances even as the linguistic identity of the tone remains constant. In this case, we would also expect vowel quality to have little effect on tone perception. Cross-linguistic comparisons have revealed that sensitivity to spectral shape is language specific. For example, Kreiman *et al.* (2010) showed that listeners who are native speakers of Gujarati, a language for which voice quality is contrastive, show heightened sensitivity to differences in harmonic amplitude relative to listeners from English and Thai, a language with lexical tone but no phonation contrast, language backgrounds. Mandarin listeners also show greater sensitivity to harmonic amplitude differences than English listeners, which could also be related to the role that voice quality plays in Mandarin, i.e., the tendency for creaky voice to cue the low tone (Kreiman and Garratt, 2010). These results indicate that perceptual attunement to spectral properties might depend both on the function of pitch and voice quality in cueing phonological contrast. While Mandarin listeners show heightened language-specific sensitivity to spectral shape, there are, in the general case, at least two possible reasons why vowel quality might not have any effect on tone perception in spite of the relationship between harmonic structure and pitch perception. As we review below, there is a general lack of concern about potential effects of vowels on tone perception in the existing literature. Much of the focus on spectral shape as related to pitch perception is on voice quality, acoustic measures of which typically attempt to correct for the influence of vocal tract resonance (e.g., Iseli *et al.*, 2007). We gather from this that one of the two possible explanations given above is (implicitly) assumed by researchers investigating tone perception.

It is also possible that due to coarticulation vowel quality might systematically influence tone perception. In the general case of coarticulation, coproduced speech gestures, such as those actuating distinct vowels, involve only a partial overlap in time. For example, in a sequence of vowels, V_1CV_2, where V represents vowels and C represents a consonant, there is anticipatory coarticulation such that the identity of V_2 influences the formant transition of V_1 into C (Ohman, 1966); there is also carryover articulation whereby V_1 influences the formant values at the onset of V_2. The direction and degree of coarticulation varies from language to language, and listeners have language-specific perceptual abilities that allow them to partially compensate in perception for the effect that one gesture has on another (Beddor *et al.*, 2002). Sequential articulation of tones also shows coarticulatory effects, which are similarly compensated for in perception (Xu, 1994). The coproduction of tones with

https://doi.org/10.1121/10.0001103

vowels constitutes a special case in that it does not involve sequential transitions in the same way—tones and vowels are either completely simultaneous or very nearly so (Gao, 2009; Xu, 2009). Tones are always coproduced with some other phonologically contrastive segment and are most acoustically salient when coproduced with a vowel or sonorant consonant (Zhang, 2002). This means that listeners in lexical tone languages do not typically experience a tone in the absence of a vowel or some other supralaryngeal constriction influencing vocal tract resonance. Although tone-vowel coproduction is a somewhat special case, there are two types of coarticulatory influences that could lead to vowel-specific tone perception.

One way that vowels could influence tone perception is through intrinsic f_0. Vowel quality has small (~11–15 Hz) but systematic effects on f_0—in general, high vowels (e.g., /i/, /u/) have higher f_0 than low vowels (e.g., /a/), an effect referred to as intrinsic f_0 (Whalen and Levitt, 1995). Intrinsic f_0 has a plausible physiological basis in shared musculature involved in modulation of tongue body position and vocal fold tension (Whalen *et al.*, 1999), which may account for its wide prevalence across languages and its suppression at the low end of the pitch range, although it seems that it can also be suppressed altogether in at least some languages (Connell, 2002). Mandarin is a typical language with respect to intrinsic f_0 in that the effect is observed outside of the low end of the pitch range (Shi and Zhang, 1987). With regard to how vowel quality might influence tone perception, the higher intrinsic f_0 for high vowels relative to low vowels could potentially bias listeners toward perceiving high tones when they are coproduced with high vowels. An assumption associated with this hypothesis is that listeners are unable to entirely compensate in perception for the effect of intrinsic f_0 on pitch modulation associated with tone. Whereas there is evidence that variation in f_0 can be interpreted as information about vowel quality (Hirahara and Kato, 1992), compensation for intrinsic f_0 may be incomplete, leading to residual effects of vowel quality on tone perception. The prediction is that tones that start high would be recognized faster on high vowels while tones that start low would be recognized faster when coproduced with low vowels.

There is another way in which physiological (mechanical) linkages between tongue body positioning and pitch modulation could induce an effect of vowel quality on tone perception. In Mandarin, tongue body height has been shown to vary systematically with tone (Erickson *et al.*, 2004; Hoole and Hu, 2004; Shaw *et al.*, 2016). Shaw *et al.* (2016) reported small but reliable effects of tone on tongue body height, some of which are consistent with the physiological explanation given for intrinsic f_0. For /a/, the position of the tongue body at the vowel target was lower when coproduced with tones that begin low (the low tone and the rising tone) than when coproduced with tones that begin high (the high tone and the falling tone). For /i/, the effects of tone on tongue body height went in the opposite direction—the tongue body was higher for tones that begin low than for tones that begin high. The effects of tone on /a/ are consistent with the "tongue-pull" hypothesis, whereby lowering of the larynx to slacken the vocal folds for low tones exerts a downward force on the jaw through the tissue connecting the laryngeal complex to the hyoid bone (Honda, 1995). In contrast, the effect of tone on /i/ may be physiologically arbitrary—it is in the opposite direction of the tongue pull hypothesis and also inconsistent with alternative laryngeal mechanisms for lowering pitch (e.g., Moisik *et al.*, 2014). For /u/, there were no significant effects of tone on tongue position at all. Together, these results indicate that there could be information about Mandarin tone contrasts in vocal tract resonances (in addition to pitch) at least for some vowels. Additionally, information about tone conveyed through tone-conditioned vocal tract shapes could be available to listeners earlier in time than other cues to dynamic tones such as f_0 turning point or creaky voice. The prediction is that tones will be recognized faster when coproduced with /a/ or /i/ than when coproduced with /u/.

To summarize, the effect that vowel quality has on tone perception is not currently known, but we have formulated a number of competing theoretical hypotheses, including some that predict no effect and some that predict specific conditions under which vowel quality influences tone perception. Table I summarizes the predictions.

B. Past work on the timecourse of tone perception in Mandarin

There has been a substantial body of work on human pitch perception by native speakers of tone languages and by speakers of non-tone languages, including on the

TABLE I. A summary of hypotheses and associated predictions for how vowel quality might influence the timecourse of tone perception.

	Hypothesis	Prediction
H_{0a}	Listeners effectively compensate for the effect of vowel quality on all harmonics	No effect of vowel quality on tone perception
H_{0b}	Listeners ignore higher harmonics that may interact with vowel resonance, focusing instead on (0)f_0 to perceive linguistic tone	No effect of vowel quality on tone perception; listeners are better at perceiving pitch in non-speech materials than in speech
H_1	Listeners fail to compensate for the effect of vowel on f_0	Tones that start high are recognized earlier on high vowels (than on low vowels); tones that start low are recognized earlier when coproduced with low vowels (than with high vowels)
H_2	Coarticulatory effects of tone on tongue body height enhance tone perception	Tones will be recognized earlier when coproduced with /a/ and /i/ than when coproduced with /u/

timecourse of tone recognition by Mandarin listeners, the group of participants relevant for this study. Some of this research has compared vowel and tone perception. For example, in a lexical decision task, vowel mismatches were more disruptive than tone mismatches in a neutral context, while this vowel-advantage was reversed by embedding words in a more predictive context (Liu and Samuel, 2007; Ye and Connine, 1999). This suggests that the relative importance of tone information can interact with context. Research on event related potentials (ERP) has revealed dissociations in neural responses to vowels and tones (Hu *et al.*, 2012).

Another key methodology for investigating the timecourse of tone recognition has been the visual world paradigm. In this paradigm, listeners hear a word and select a matching picture or orthographic representation on a computer monitor. The listeners' eye movements are tracked as they complete the task to measure the timecourse of word recognition as they select between the displayed options. This paradigm was used by Malins and Joanisse (2010), who found that looks toward the target picture diverged from competitor pictures when disambiguating acoustic information became available in the signal regardless of whether it was attributable to a vowel or a tone. This demonstrated that vowels and tones mutually constrained lexical access. In that study, results were averaged across stimulus items containing different vowels and tones. Shen *et al.* (2013) reported similar results in two experiments using the visual world paradigm, one with pictures and another with characters. Their stimuli manipulated the f_0 of target items while holding constant all other acoustic aspects of the materials, including (presumably) higher level harmonics. Manipulations of f_0 height, particularly at the syllable onset and offset had significant effects on looks to target, indicating that lexical tone perception by Mandarin listeners is an incremental process sensitive to pitch heights at critical points in the syllable. This conclusion meshes well with the theoretical proposal of House (1996) that listeners in tone languages are more sensitive to f_0 at the beginning and ends of vowels as this is where disambiguating information tends to be concentrated. In another study using the visual world paradigm with characters as visual stimuli, Wiener and Ito (2015) showed that the probability of the tone given the segments influenced the timecourse of tone recognition for low frequency words. This result reveals another way, in addition to our hypotheses presented in Table I, that a vowel can influence tone perception. Wiener and Ito (2015) also reported that there was no effect of tone probability on tone recognition in the reaction time of the mouse clicks, indicating the sensitivity of eye movements as a measure of online perception of tones.[2]

The studies described above provide important insights into the timecourse of tone processing by Mandarin speakers, including that processing is incremental across the syllable (Shen *et al.*, 2013), word recognition is mutually (and simultaneously) constrained by tones and vowels (Malins and Joanisse, 2010), and non-equiprobability of tones in the lexicon also shapes recognition (Wiener and Ito, 2015). These results have since been incorporated into models of spoken word recognition that include tone (Shuai and Malins, 2017). They leave open, however, the possibility that vowel identity conditions tone recognition. For this question, it is crucial to have stimuli that bear the coarticulatory signature of naturally produced words and systematically vary tone-vowel combinations, design considerations that we have factored into the current study.

C. This study

To test our hypotheses, we require a task that will allow us to measure the time it takes to recognize a word when tone is the only discriminating feature. The visual world paradigm provides fine-grained information about the timecourse of recognition, but there are many factors related to the word itself that could affect recognition of the auditory stimulus. Word frequency, neighbourhood density, and onset density all have an effect on the timecourse of recognition (Magnuson *et al.*, 2007). In the case of Mandarin Chinese, pitch shape (i.e., static vs dynamic; Wu and Ortega-Llebaria, 2017), syllable frequency, and tonal probability (Wiener and Ito, 2015, 2016) have an influence on spoken word recognition, and stroke complexity affects visual word recognition (Liversedge *et al.*, 2014). These potential confounds make it difficult to choose minimally contrasting stimuli that are controlled for lexical statistics.

To isolate the effect of tone identity on word recognition in Mandarin Chinese and to counteract the effects of lexical statistics, we compared participants' fixations to the target word in two conditions. In a control condition, the target word was presented on the screen with three distractors that were distinguishable from the target on the basis of the initial consonant (e.g., /ta55/ vs /t^hu214/, /pa35/, and /t^hi35/). This provides a baseline of the timecourse of recognition for each target token in conditions where the listener is not required to make a decision based only on tonal identity. In the experimental condition, the visual target was presented with competitors having the same onset consonant and the same vowel but different tones (e.g., /ta55/ vs /ta35/, /ta214/, and /ta51/). In that case, the only information that the listeners could use to decide on the correct visually presented word is the tone. Since the time taken to recognize each word may vary due to extraneous factors, such as lexical statistics, the measure of interest is the difference between looks to target in the baseline condition and looks to target in the experimental condition. We expect word recognition to be slower in the experimental condition than in the control condition, but the question of interest is whether the difference varies as a function of vowel within each tonal category.

To further mitigate against confounding effects of stroke order and lexical statistics on recognition of the orthographically presented response options, we provided listeners with a preview of the visual stimuli before the trial began and enforced a neutral gaze requirement. At the beginning of each trial before hearing the auditory stimulus,

https://doi.org/10.1121/10.0001103

listeners first looked at the four response options and clicked the mouse when they were ready to proceed. This was the preview phase of the trial. The neutral gaze requirement forced the eyes to fixate on a crosshair in the center of the screen. Fixation on the crosshair triggered presentation of the auditory stimulus. Besides allowing time to process all items on the screen before auditory stimulus presentation, the preview window may have further focused listener attention on relevant phonetic aspects of the signal. In control conditions, the preview window revealed to participants that the initial consonant differentiates competitors. In test conditions, the preview window revealed that only tone differentiates competitors.

Finally, our design featured the re-use of the same auditory and visual stimuli in both target and control trials. The auditory stimulus tokens were repeated four times across the course of the experiment (twice in the control condition and twice in the experimental condition), and the visual stimuli were repeated 96 times (24 as target and 72 as distractor/competitor). Repetition of items, together with the preview window, plays into our general strategy of minimizing lexical effects so that we can focus on the role that vowel quality may play in the uptake of acoustic information about tone.

II. METHODS

A. Subjects

The sample of listeners consisted of 16 native speakers of Mandarin Chinese (2 male; mean age = 24 yr; standard deviation = 6 yr; age range, 19–47 yr) from the Western Sydney University community. The speakers grew up in Northern China as far south as Wuhan and moved to Australia as adults after the age of 18. Two additional participants were tested, but their data were discarded due to technical problems with data acquisition.

B. Materials

The stimuli consisted of 28 syllables from Mandarin (see Table II). The syllables were all of the shape CV, where C indicates an onset consonant and V indicates a monophthong vowel. The vowels were drawn from the set /a/, /i/, /u/. These three vowels were combined with all four Mandarin lexical tones and three onset consonants: /p/, /t/, and /t^h/.

TABLE II. The 28 Chinese characters used as stimuli in the experiment, together with IPA transcriptions for the segments, grouped by vowel and tone. The tone numbers provided in the left column are the standard descriptors for Mandarin tones—correspondence to IPA tone numbers and labels is as follows: tone 1 = 55 high, tone 2 = 35 rising, tone 3 = 21(4) low (falling rising), and tone 4 = 51 falling.

Tone	Vowel		
	/a/	/i/	/u/
Tone 1	搭 /ta/ 八 /pa/	低 /ti/ 逼 /pi/ 梯 /t^hi/	嘟 /tu/ 突 /t^hu/
Tone 2	达 /ta/ 拔 /pa/	敌 /ti/ 鼻 /pi/ 提 /t^hi/	读 /tu/ 图 /t^hu/
Tone 3	打 /ta/ 把 /pa/	底 /ti/ 比 /pi/ 体 /t^hi/	赌 /tu/ 土 /t^hu/
Tone 4	大 /ta/ 爸 /pa/	地 /ti/ 必 /pi/ 替 /t^hi/	度 /tu/ 兔 /t^hu/

Like English, Mandarin consonants contrast in aspiration; /t/ is differentiated from /t^h/ by voice onset time (VOT), which is short for /t/ and long for /t^h/ (Chen *et al.*, 2007; Lisker and Abramson, 1964). The stimulus recordings were sourced from an existing corpus at the MARCS Institute for Brain, Behavior and Development, Western Sydney University, and are reported in Xu Rattanasone *et al.* (2014). The full set of tone × vowel combinations was only available for the /tV/ context; the /pV/ and /t^hV/ contexts were included as filler items only. The stimuli are presented in Table II.

The syllables were recorded by three female native speakers of Mandarin Chinese from northern China. A total of 6 tokens of each syllable were included in the experiment with one to three tokens per speaker, 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 10 Hz sixth-order elliptical high-pass filter was applied to remove the direct current (DC) component of the signal.

For each sound file, acoustic measurements for f_0, F_1, F_2, and F_3 were sampled every 16.67 ms using Praat software (Boersma and Weenink, 2013). The formants were calculated using linear predictive coding (LPC), the Burg method. The window length was 25 ms, pre-emphasis was applied from 50 Hz, and the parameters for maximum number of formants and the maximum formant frequency were optimized on a token-by-token basis by visual inspection of the spectrogram. Here, we report the mean acoustic measurements for the critical /tV/ syllables only. The f_0 contours for each tone × vowel are presented in Fig. 1. To calculate mean formant values, we used the method of sampling the formants by Zee (1980) and then averaged them across the vowel. The mean F_1, F_2, and F_3 values are presented in Table III.

C. Procedure

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

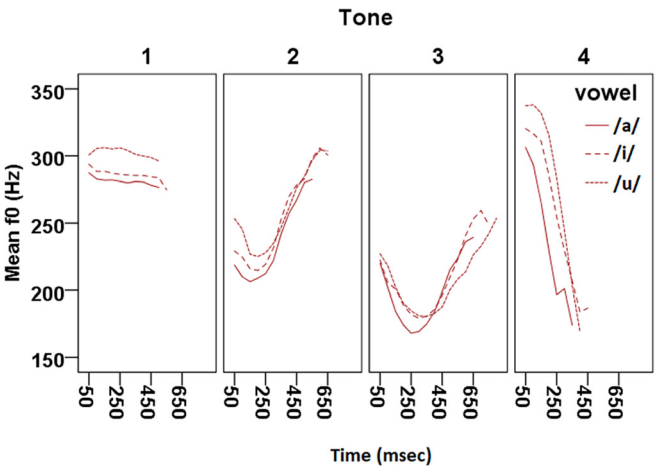

FIG. 1. (Color online) Mean *f*0 contours for target stimuli (/t/ context only) in each tonal context, split by vowel.

TABLE III. Formant measurements, in Hz, averaged across the vowel and across the three speakers, for each vowel × tone combination. Standard deviations are in parentheses.

Vowel	Tone	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)
/a/	1 (high)	903 (103)	1635 (99)	2971 (22)
	2 (rising)	972 (34)	1646 (90)	3084 (246)
	3 (low)	984 (26)	1588 (31)	2997 (346)
	4 (falling)	1012 (19)	1617 (72)	2932 (68)
/i/	1 (high)	481 (9)	2723 (190)	3523 (110)
	2 (rising)	472 (17)	2723 (305)	3567 (191)
	3 (low)	497 (20)	2715 (202)	3578 (216)
	4 (falling)	537 (49)	2710 (190)	3436 (147)
/u/	1 (high)	452 (24)	1017 (68)	2550 (273)
	2 (rising)	505 (33)	892 (66)	2741 (106)
	3 (low)	445 (14)	948 (23)	2713 (196)
	4 (falling)	495 (60)	952 (125)	2619 (311)

a chin rest located 50 cm from the monitor. Auditory stimuli were played over loudspeakers located next to the monitor. A Tobii-x120 eye tracker (Danderyd, Sweden) 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 (Fig. 2, left). 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 rectangle appeared around the crosshair (Fig. 2, middle). When continuous gaze duration to the crosshair reached 200 ms, the crosshair and red rectangle disappeared and the auditory stimulus played (Fig. 2, right). This procedure allowed participants to preview words in the four quadrants of the screen and ensured that they were fixating on the center of the screen at the start of the auditory stimulus.

The auditory and visual stimuli were organized into two within-subjects conditions, a test condition and a control condition. In the test condition, the target word was shown on the screen (in one of the four quadrants) along with competitor words differing only in tone (see Fig. 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. Deviations from the baseline word-recognition time in the test condition could

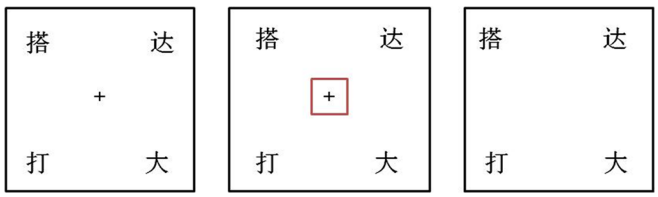

FIG. 2. (Color online) Schematic depiction of an experimental trial (test condition). Participants clicked on the crosshair (left), fixated on the cross hair (middle), and clicked on the word that they heard (right).

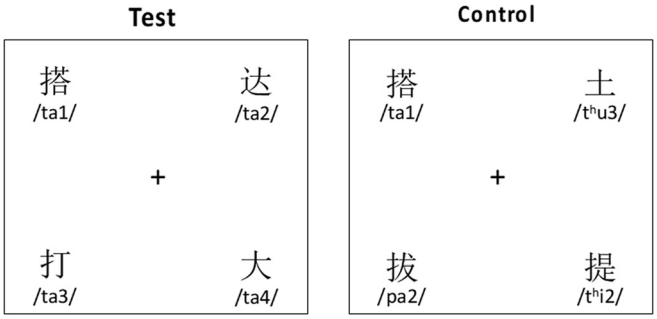

FIG. 3. Examples of test (left) and control (right) conditions. In these examples, 搭 /ta1/ is the target. In the test condition, competitors differ only in tone. In the control condition, competitors differ in both the onset consonant *and* either vowel or tone. The Romanized transcriptions are included for illustrative purposes only. They were not presented to participants. The numbers in slashes refer to the Mandarin tones using standard descriptors—correspondence to International Phonetic Alphabet (IPA) tone numbers and labels is as follows: tone 1 = 55 "high," tone 2 = 35 "rising," tone 3 = 21(4) "low (falling rising)," and tone 4 = 51 "falling."

then be attributed to the role of tone in word recognition. The presentation of all stimuli, including control and test conditions, was fully randomized for each participant.

Trials containing each auditory token occurred twice in the test condition and twice in the control condition. The visual stimuli were counterbalanced so that each character appeared as the target three times in each screen quadrant for control trials and three times in each quadrant for test trials. The visual stimuli also appeared as distractors 9 times in each quadrant for test trials and between 4 and 12 times per quadrant for control trials. Each sound file was played 4 times for a total of 672 trials (28 syllables × 6 tokens × 4 repetitions), which were presented in random order. Participants were given a break halfway through the experiment.

D. 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 role of tone in picking out the target in the test condition (cf. onset differences in the control). To compute target fixation, we coded eye fixations for the 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 were 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, following Barr (2008). The elogit transformation converts fixation proportions to a continuous scale without upper or 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 regression model (see Sec. III, Results).

We determined the time window for analysis by plotting the elogit across all conditions as a function of time and

https://doi.org/10.1121/10.0001103

identifying (1) a sharp increase in looks toward the target and (2) a plateau in looks toward the target. This was done across conditions to eliminate hypothesis-based bias in window selection (Barr, 2008). The selected analysis window begins at 300 ms and ends at 800 ms. Within this analysis window, looks to the target increase linearly as a function of time. 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 (Matin *et al.*, 1993). The duration of the window is greater than the duration of the stimuli, which ranged between 250 and 500 ms, and therefore can reflect looks driven by phonologically relevant information distributed across the word.

III. RESULTS

A. Visualization

Figure 4 shows target fixation (elogit transformed) across conditions (separate lines) for each combination of tone and vowel for items following the consonant /t/. For all tones, there were fewer target fixations in the test condition than in the control condition, indicating that, as expected, the timecourse of syllable recognition is slowed when there are only lexical tone competitors. The degree to which tone competitors impact word recognition beyond dissimilar competitors is represented in the differences between the lines (control vs test) in each panel of Fig. 4. The closer the lines the faster the tone is recognized. Figure 5 collapses across time to show the mean difference (between target and control conditions) in target fixation for each tone and vowel combination. Shorter bars indicate faster tone recognition.

It is clear, first, that the timecourse of tone recognition is not uniform across vowels. Most tones tend to be recognized faster when they are coproduced with /a/ or /i/ than when coproduced with /u/. This difference is particularly robust for the high and rising tones. Across tones, the difference between the lines tends to be larger for /u/ than for /a/ or /i/.

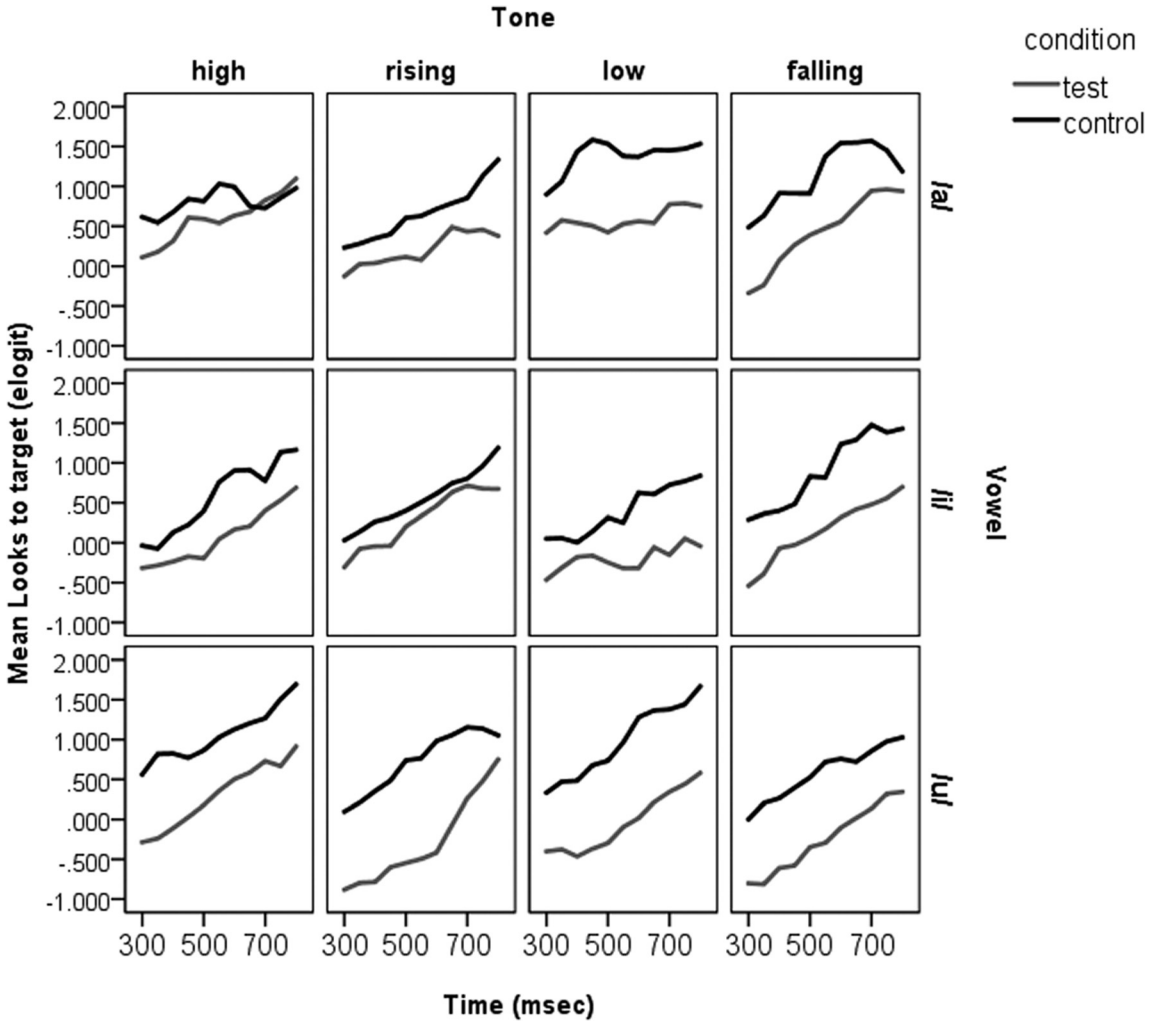

FIG. 4. Target fixation (y axis) by time (x axis) for each tone (columns) and vowel (rows) combination. The lines show different conditions: black lines show the control condition, in which competitors differed from targets in the onset consonant (as well as the vowel or tone); gray lines show test trials, in which competitors differed from targets only in tone. The difference between lines is the effect of tone perception on word recognition. When control and test lines are close together, tone perception is fast; greater separation between control and test indicates slower tone recognition.

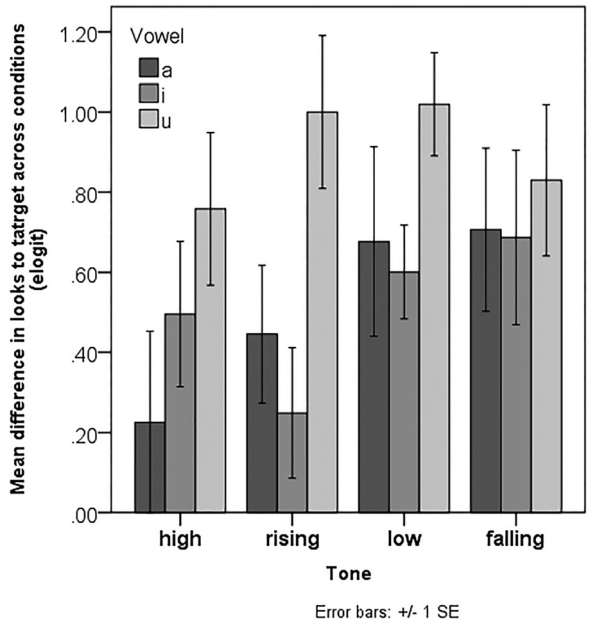

FIG. 5. Mean difference in participants' target fixation (y axis) between test and control conditions in the 300–800 ms time window by tone (x axis) and vowel (separate bars) combinations. Error bars represent standard error of the mean. Shorter bars indicate faster tone recognition.

B. Statistical models

To evaluate the statistical significance of differences in target fixation observed across tones and vowels, we fit a series of nested linear mixed effects regression models to the elogit-transformed data with crossed random effects for subject and item. Our choice of a linear model was motivated both by the linear increase in our dependent variable across the analysis window and the categorical nature of our predictors, cf. analysis of visual world data using continuous predictors (e.g., Porretta *et al.*, 2016).

Our baseline model contained the fixed factor of time, expressed in ms, random intercepts for both subjects and items and by subject and by item random slopes for time. The random slopes were included to capture subject and item differences in how quickly gaze converged on the target. To the baseline model, we added condition (control, test), tone (high, rising, low, falling), and vowel (a, i, u), as well as the interactions between tone and condition, between vowel and condition, and the three-way interaction between tone, vowel, and condition. Vowel was treatment coded with /u/ as the intercept. We chose /u/ for the intercept because of work in speech production indicating that the position of the tongue varies across tones for /a/ and /i/ but remains stable for /u/ (Shaw *et al.*, 2016). Tone was also treatment coded with the high tone as the intercept. Condition was coded numerically with the control as 0.5 and the test condition as −0.5. This dictates that the control condition will be rendered as the intercept. We also explored models including the log character frequency of target items, but this was eliminated from the final models because its inclusion did not significantly improve model fit or interact with other factors. Models were coded in *R* (*R* Development Core Team, 2006) using the *lme4* package (Bates *et al.*, 2015).

Table IV summarizes the comparison of nested models. Relative to the baseline model with only time as a fixed factor, adding condition led to substantial improvement as reflected in the large increase in the log likelihood (logLik) of the data. Adding the tone * condition interaction term[3] also resulted in a statistically significant improvement to the model. This improvement also comes with a substantial increase in model complexity—the degrees of freedom increase from 10 to 17 because of the added coefficients for each combination of tone and condition. The increase in model complexity is justified by improved performance as indicated by the decrease in the Akaike information criterion (AIC) from 8862 to 8833. The addition of the main effect of vowel to the model yields modest additional improvement—the Chi-squared χ^2 test is significant ($p < 0.01$), and there is a decrease in AIC. A much larger improvement is achieved by adding the full three-way interaction between vowel and tone and condition. Although the degrees of freedom increase to 32, this comes with a substantial decrease in AIC.

To explore the interaction between vowel * tone * condition, we fit another set of models to each tone separately. For each tone, we explored whether the interaction term between vowel and condition led to significant improvement over a model containing only the main effects. The tone-specific models had the same random effects structure as the models summarized in Table IV.

Tone-specific model comparisons are summarized in Table V. The interaction between vowel and condition is significant for three out of the four tones—high, rising, and low but not the falling tone. The direction of the effects is indicated by the coefficients in the regression models. Since the control condition defines the intercept, a negative β coefficient is expected for all tones and indicates that there were fewer looks to the target in the test condition (i.e., when there were tone competitors) than in the control condition.

TABLE IV. Comparison of the models fit to target fixations (elogit).

Model of looks to target (elogit)	Degrees of freedom (df)	Akaike information criterion (AIC)	Log likelihood (logLik)	Chi squared statistic (Chisq)	Pr(<Chisq)
Time + (1 + time|subject) + (1 + time|item)	9	9473	−4728		
Condition + time + (1 + time|subject) + (1 + time|item)	10	8862	−4421	612.79	<0.0001
Tone * condition + time + (1 + time|subject) + (1 + time|item)	17	8833	−4399	43.66	<0.0001
Vowel + tone * condition + time + (1 + time|subject) + (1 + time|item)	19	8827	−4394	9.88	<0.01
Vowel * tone * condition + time + (1 + time|subject) + (1 + time|item)	32	8711	−4323	142.22	<0.0001

https://doi.org/10.1121/10.0001103

TABLE V. Comparison of tone-specific models fit to target fixations (elogit).

Model of looks to target (elogit)		Df	AIC	logLik	Chisq	Pr(<Chisq)
Tone 1 (high)	Vowel + condition	12	2190	−1083		
	Vowel * condition	14	2162	−1067	32.10	<0.0001
Tone 2 (rising)	Vowel + condition	12	2145	−1060		
	Vowel * condition	14	2101	−1036	48.20	<0.0001
Tone 3 (low)	Vowel + condition	12	2099	−1037		
	Vowel * condition	14	2073	−1022	29.94	<0.0001
Tone 4 (falling)	Vowel + condition	12	2316	−1146		
	Vowel * condition	14	2319	−1146	0.21	0.9026

A positive condition * vowel interaction term for /a/ or /i/ indicates that for these vowels the effect of condition was not as great as for the intercept vowel /u/. In other words, vowel modulates (and more specifically in this example, attenuates) the effect of condition. Certain vowels bring tone perception in the presence of tone competitors closer to the recognition behavior in the control condition where tone competitors are absent. Certain vowels facilitate tone perception. For tone 1 (high), the interaction terms for both /a/ ($b = 0.55$; $t = 5.68$) and /i/ ($b = 0.31$; $t = 3.29$) were positive and statistically significant with the effect for /a/ stronger than /i/. This indicates that the vowel environments /a/ and /i/ facilitate high tone perception. For tone 2 (rising), again, both interaction terms for /a/ ($b = 0.52$; $t = 5.60$) and /i/ ($b = 0.63$; $t = 7.03$) were positive and significant, this time with /i/ stronger than /a/. For tone 3 (low), the coefficients for both /a/ and /i/ were positive, but only /i/ was statistically significant ($b = 0.46$; $t = 5.41$, cf. for /a/, $b = 0.14$; $t = 1.41$). As mentioned above, the interaction was not significant for tone 4 (falling), indicating that the effect of condition is uniform across vowels for this tone.

IV. DISCUSSION

Using a novel variant of the printed word visual world paradigm, we investigated how vowel coproduction influences the timecourse of tone perception in Mandarin Chinese. We formulated multiple competing hypotheses, including two versions of the null hypothesis which predict that there will be no effect of vowel on tone recognition, as well as two alternatives which predict specific vowel-tone interactions. Our results indicated clear effects of vowel coproduction on the timecourse of tone recognition. For three of the four lexical tones, tone recognition was conditioned by vowel coproduction. In light of these results, we return to the predictions made by our hypotheses.

Our first hypothesis followed from the observation that some vowels condition a higher f_0 than others. Specifically, intrinsic f_0 is higher for /i/ and /u/ than for /a/, and this is true both within Mandarin, generally, and for the materials used in this study. If listeners fail to completely compensate for the effect of vowel quality on f_0, then they may attribute higher f_0 in /u/ and /i/ to a high tone. This would result in earlier looks to targets when tone 1 (high) and tone 4 (falling) are produced with /u/ and /i/ and earlier looks to targets when tone 3 (low) and tone 2 (rising) are produced with /a/. This is plausible since artificial manipulations of f_0 of a similar magnitude as intrinsic f_0 have been shown to influence tone perception in this type of paradigm (Shen *et al.*, 2013). However, our specific results did not corroborate these predictions. Tones that start high did not pattern together; there was no effect of vowel on the falling tone; and for the high tone, it was actually /a/, which has low intrinsic f_0, that facilitated speeded recognition. The tones that start low, e.g., tone 2 (rising) and tone 3 (low), also did not pattern together. The rising tone was recognized faster when produced with /a/ than when produced with /u/, which is the one result that is partially consistent with this hypothesis, but it was /i/ that best facilitated recognition of the rising tone. For the low tone as well, /i/ facilitated faster recognition than /u/ or /a/. Overall, then, the specific pattern of results does not support H_1 very well. The effect of vowel on tone recognition in Mandarin does not seem to follow from a misattribution of intrinsic f_0 to tone.

The alternative hypothesis was that coarticulatory effects of tone on tongue body height could enhance tone perception. Research on the articulation of vowels in different contexts has shown tone influences tongue body height, jaw height, and the resulting acoustics, particularly F_1 (Erickson *et al.*, 2004). However, these effects are not uniform across vowels. Shaw *et al.* (2016) found that only the vowels /a/ and /i/ varied in articulation as a function of tone coproduction; articulation of /u/ was entirely unaffected by tone. One possible explanation for the stability of /u/ involves the role of the jaw in tone-vowel coproduction. Jaw movement in speech involves both rotation around a hinge, the temporomandibular joint, and vertical/horizontal translation (Edwards and Harris, 1990); notably, rotation can be controlled independently of translation (Ostry and Munhall, 1994). The pattern of tone influence on tongue position for vowels is potentially driven by the rotation component of jaw movement. This could explain why, for example, tone does not influence the posterior portion of the tongue; Shaw *et al.* (2016) reported that tongue dorsum position was stable across the four Mandarin tones. Since the tongue dorsum is closer to the temporomandible joint than the tongue body or tongue tip, rotation of the jaw around this hinge will have less effect on tongue dorsum position than on more anterior portions of the tongue. Similarly, the stability of the tongue body and tongue tip across tone for /u/ could be related to

the posterior vowel target. That is, for /u/ the whole tongue moves toward the temporomandible joint, which could minimize influences of jaw rotation on tongue position. Consistent with this account, other vowels in the vicinity of /u/, i.e., /ou/, /uo/, appear to be similarly resistant to influences of tone (Shaw and Chen, 2019). The resistance of /u/ to tone-based coarticulatory effects has the consequence that there is no information about tone contained in the vocal tract resonance when tones are coproduced with /u/.

Consistent with this second hypothesis, we found that there were perceptual advantages in tone recognition when tones were coproduced with /a/ and /i/ as opposed to /u/. Specifically, for three out of four tones, recognition was faster when the tones were coproduced with /a/ and /i/, the vowels that are also subject to tone-conditioned articulatory variation. This result is largely consistent with the hypothesis that the influence of tone on vocal tract resonance is perceived as information about tone.

In speech production, the Mandarin tones that start high, the high (tone 1) and the falling (tone 4) tones, and tones that start low, the low (tone 3) and rising (tone 2) tones, pattern together in their influence on tongue body height. If coarticulatory influences of tone production on tongue body height serve as cues to tone identity, then we would expect facilitatory effects of vowel coproduction to pattern together across the two tones that begin low and across the two tones that begin high. Consistent with this expectation, the low and the rising tones patterned together in perception. Recognition of these tones was speeded by coproduction with /a/ and /i/ relative to /u/. In contrast, however, tones that begin high did not pattern together. The expected advantages of vowel coproduction were observed for the high tone with /a/ and /i/ recognized faster than /u/ but were not observed in the falling tone condition. For the falling tone, there was no effect of vowel on the timecourse of tone recognition.

It is not entirely clear why differences in vowel coproduction on tone recognition were not observed for the falling tone. One possibility involves the way that vocal tract resonance selectively amplifies harmonics. To the extent that harmonics cue pitch in speech perception (cf. the hypothesis from the introduction that they do not). the amplitude of various harmonics will be relevant. Naturally, the amplitude of harmonics will be conditioned in part by vowel quality. Harmonics that happen to coincide in frequency with a vowel resonance will be amplified while others will be damped. Amplification of individual harmonics may increase the acuity of pitch perception, effectively increasing the relative amplitude of aspects of the signal relevant for recovering pitch. The resonant properties of the vocal tract and the harmonics of vocal fold vibration are both dynamic as both change over time with the movement of the articulators, but they are largely independent. For example, it is possible to observe falling pitch, characterized by a decrease in all harmonics over time concurrently with rising F_2, as the tongue body moves forward in the vocal tract decreasing resonant frequencies of the front cavity that contribute to F_2. In a case such as this, the dynamics of harmonics and resonances can be poorly aligned, resulting in low amplitude harmonics. In contrast, they could also be aligned more closely such that changes in vocal tract resonance over time function to enhance rather than dampen particular harmonics relevant to the perception of tone. This type of acoustical interaction between vocal tract resonance and source harmonics could potentially influence tone perception in a way that is orthogonal to our main hypotheses, causing our predictions to go astray in particular acoustical situations.

We now examine in greater detail the possibility of overlap between harmonics and vowel resonance in the specific case of tone 4 (falling tone) where we did not see the anticipated effect of vowel on tone recognition. Figure 6 plots the estimated harmonics for the falling tone in Mandarin as coproduced with three vowels, /a/ (top), /i/ (middle), and /u/ (bottom), for each of the stimulus tokens in

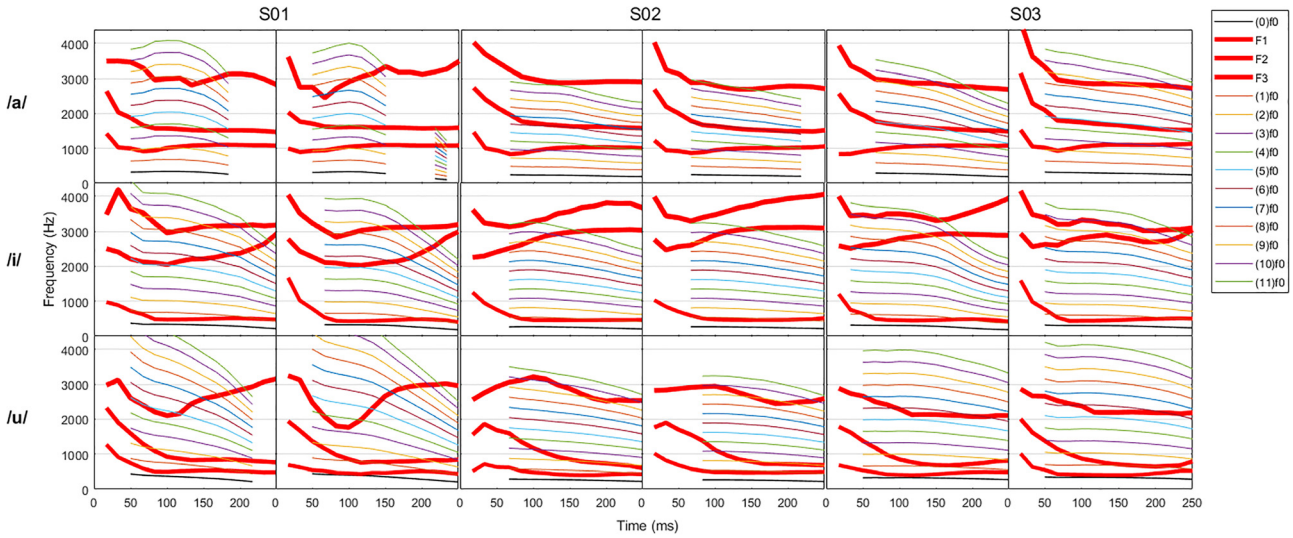

FIG. 6. (Color online) f_0 and estimated harmonics for the falling tone over time in the three vowel contexts. The thick red lines show F_1, F_2, and F_3, which overlap with harmonics, particularly (2)f_0, just in the /u/ vowel context.

https://doi.org/10.1121/10.0001103

the experiment. These include two repetitions of each item from each of three speakers, labelled *S*01, *S*02, *S*03. Figure 6 shows the first 250 ms of each token, which includes a period of silence for the initial stop consonants. The harmonics, $1f_0$–$11f_0$, were estimated to be integer multiples of $(0)f_0$. As a proxy for vowel resonance for the sake of illustration, we have overlaid the first three formants, F_1, F_2, F_3, of each vowel as a red line, including LPC interpolation across the initial consonants (for methods, see Sec. II B). The thickness of the red lines is related to formant bandwidth.[4] For this illustration we abstract away from the important issue of how harmonics impact formant estimation (see, e.g., Shadle *et al.*, 2016). The key observation is that the degree to which the formants overlap harmonics varies with vowel quality. For /a/ (Fig. 6, top), there is some overlap at the start of the vowel between F_2 and $5f_0$. After 150 ms, however, $5f_0$ and F_2 diverge as $5f_0$ lowers at a faster rate than F_2. The other formants show less consistent patterns. Early in the vowel F_1 changes in the opposite direction of the harmonics, rising as the harmonics fall. For /i/, there is even less overlap (than for /a/) between F_2 and any harmonic— F_2 *rises* as the harmonics fall. The situation is similar for F_3. F_1 shows a sharp initial fall and then a plateau but overlaps early in the vowel with $1f_0$, particularly for the two tokens of *S*02. The vowel with the greatest overlap between lower formants and the harmonics is /u/. Early in the vowel, both F_1 and F_2 fall. The harmonics fall as well, and for many tokens there are periods of substantial overlap between the decrease in formant frequencies and the decrease in harmonics. In this scenario, by virtue of overlap with a vowel resonance, lower harmonics, e.g., $2f_0$, will have increased amplitude in /u/ relative to /a/ and particularly /i/. All else equal, increased amplitude of a particular harmonic can be expected to improve pitch perception. We speculate that harmonic amplification may have made it easier to perceive the falling tone with /u/ and thus balanced out the other advantages that coproduction with /a/ and /i/ were expected to have, according to our second hypothesis, H_2 (Table I). If our speculation is correct, then the null effect of vowel on the falling tone (tone 4) could be due to two factors related to vowel coproduction that cancel each other out in this particular environment.

With respect to our main hypotheses, we found that much of the data are consistent with the idea that vocal tract resonances themselves contain information relevant to identifying lexical tones. In this way, tone-conditioned coarticulation provides a cue to tone identity much in the way that coarticulation is known to cue perception of consonants and vowels (Beddor *et al.*, 2002; Beddor *et al.*, 2013; Benguerel and McFadden, 1989; Mills, 1980; Nittrouer and Studdert-Kennedy, 1987). This is likely because the physiology of lexical tone production in Mandarin exerts a coarticulatory influence on tongue body height at least for some vowels. Listeners appear to be sensitive to this variation and use it as information about lexical tone.

This main result adds to the large number of factors already known to influence tone perception, including phonetic dimensions of the stimuli beyond pitch (e.g., duration, intensity, and voice quality), lexical factors, and the language experience of the listener, which may interact with both lexical and phonetic factors. Our study did not investigate language experience. The stimuli were ostensibly Standard Mandarin, produced by speakers from Beijing; the listeners came from a slightly wider range of dialect backgrounds in mainland China. Perception of Standard Mandarin tones is known to be influenced by listener knowledge of other language varieties, including both quite distinct languages such as Cantonese (Zhang *et al.*, 2012; Wiener and Ito, 2016), as well as varieties that are more similar to Standard Mandarin such as Jinan Mandarin (Wu *et al.*, 2019). The degree to which vowel-conditioned tone processing generalizes across populations with differing degrees of language experience is an interesting question for future research. Some aspects of tone-conditioned vowel variation, such as lower jaw and tongue body for /a/ produced with low tones, appear to follow from physiological constraints on speech production while others, such as higher tongue body for /i/ produced with low tones, may be language specific. Although we observed vowel-conditioned tone perception for both of these vowels in our sample of listeners, it is possible that the facilitatory effects of /a/ would generalize across a wider range of listeners than the facilitatory effects of /i/.

We have also raised the idea here that "fortuitous" enhancement of harmonics by vocal tract resonances can function as a possible additional factor relevant to explaining variation in the timecourse of tone perception across vowels. In natural speech, lexical tones are coproduced with different vowels and consonants in line with the unique speech anatomy and physiology of individual talkers. The combination of these factors leads to scenarios in which the change over time of particular harmonics may follow vocal tract resonances to various degrees. While this could lead to a fortuitous enhancement of tone perception when the vocal tract resonances enhance harmonics, it may also slow tone perception when harmonics are misaligned with vocal tract resonances (and therefore dampened). The "fortuitous" enhancement account assumes that listeners are attuned to harmonics and use this information for pitch perception, cf. the hypothesis raised in the introduction that they are not. For the case of Mandarin listeners, this is plausible as past research has established that Mandarin listeners show particular sensitivity to relative differences in the amplitude of harmonics (Kreiman and Gerratt, 2010). In Mandarin, creaky voice can occur during intervals of low pitch, including at the end of the falling tone (tone 4), the beginning of the rising tone (tone 2), and the middle of the low tone (tone 3; Kuang, 2017) but only provides a perceptual cue to the low tone (tone 3; Huang, 2019). In our own materials, creaky voice occurred in 7 of the 72 target tokens, 6 times on tone 3 and 1 time on tone 4. It is unlikely, given our design, that this played any role in the vowel-conditioned patterns of tone perception, but it is possible that the role of creaky voice in Mandarin as a cue to tone 3 contributes to a language-specific attunement to higher harmonics. If so, we

might expect vowel-based enhancement effects—to the extent that this account is on the right track—to be language specific and more likely in languages that enhance pitch differences for lexical tones with voice quality cues.

An idea related to our proposal has been put forward to explain patterns in the cross-linguistic typology of tone systems (Gordon, 2001; Zhang, 2002). Zhang (2002) argues convincingly that language-specific constraints on the distribution of contour tones are related to how well particular contexts support the perceptual recoverability of tone. The survey by Gordon (2001) of 105 languages reveals a tendency for contour tones (e.g., falling or rising tones) to be restricted to certain syllable types. For each language in the survey, Gordon tabulated which syllable types permitted contour tones. A total of four syllable types were considered: open syllables with long vowels (i.e., CVV), syllables closed with a sonorant consonant (i.e., CVR), syllables closed with an obstruent (i.e., CVO), and open syllables with a short vowel (i.e., CV). Some languages in the survey ($n = 12$) did not allow any contour tones, and others allowed contour tones on all syllable types ($n = 36$). The remaining languages restricted contour tones to only CVV syllables ($n = 25$), only CVV and CVR syllables ($n = 29$), or only CVV, CVR, and CVO syllables ($n = 4$). This pattern of tone restrictions constitutes an implicational hierarchy: CVV $\gg$ CVR $\gg$ CVO $\gg$ CV, such that contour tones are not allowed at one level of the hierarchy unless they are also allowed at higher levels. Gordon argues that the phonetic basis for this cross-linguistic pattern resides in the degree to which different syllable types amplify the harmonics relevant to pitch perception. Vowels enhance harmonics to a greater degree than sonorants, and sonorants enhance harmonics to a greater degree than obstruents. The extension of this idea that we have advanced in the discussion of our data is that particular combinations of consonants and vowels may also lead to changes of vocal tract shape over time, which may amplify harmonics to varying degrees. If this is on the right track for our data, the size of such effects would be on the order of magnitude of those contributed by the coarticulatory effects of tone on tone body height. At this point, however, we do not have a method to rigorously evaluate this possibility.

The main result of this study, the finding that vowels influence the timecourse of tone perception, has a range of consequences for how we understand the phonetics of tone perception in natural language. There is by now a substantial body of work on lexical tone perception, including the timecourse of tone perception; however, much of the research ignores tone-vowel coarticulation. For example, Lee *et al.* (2008) investigated the effect of degraded acoustic stimuli on Mandarin tone perception, including one condition in which listeners were asked to identify tone based only on the first six pitch periods of the rime. Results in this condition demonstrated that recognition of tone 1 (high) and tone 2 (rise) is most impacted by truncating the syllable. This would seem to indicate that early information is more important for recognizing the high and rising tones of Mandarin than for recognizing the low and falling tones. However, our results on tone-vowel interactions cast doubt on this conclusion and others based on materials that do not control for vowel. In our data, tone 1 and tone 2 are indeed recognized early in time for vowels /a/ and /i/ as indicated by the closeness of the control and test lines in Fig. 4; but, this is not the case for tone 1 and tone 2 when they are coproduced with /u/. As it turns out, most of the tone 1 items in the study by Lee *et al.* happened to be coproduced with /i/. It is therefore not clear whether these and other results about tone processing generalize beyond the specific vowel contexts in which they were tested.

In our own study, we used two strategies to control for lexical, phonotactic, and orthographic factors that could potentially influence our interpretation of the results. Allowing participants to preview the visual stimuli minimized the influence of differences between the times taken to recognize individual characters. Factors relating to the time taken to recognize each auditorily presented word were minimized by comparing the results of our experimental trials to a control condition where the word could be recognized on the basis of the initial consonant. By isolating phonetic factors, our design complements studies investigating lexical competition (e.g., Malins and Joanisse, 2010; Shuai and Malins, 2017; Wiener and Ito, 2016). Since the design of this study is novel, there are naturally aspects that should continue to be explored. These include whether our attempts to minimize lexical, phonotactic, and orthographic effects with this paradigm were successful. Besides testing the generalization of our findings to a broader range of consonant and vowel contexts in Mandarin Chinese, future research should investigate interactions between phonetic factors, such as vowel quality, and lexical factors, such as frequency, tonal probability, and the distribution of competitors, which are also known to influence the timecourse of tone recognition.

V. CONCLUSIONS

Using a printed character version of the visual world paradigm, we investigated whether vowel quality influences the timecourse of tone recognition in Mandarin Chinese. Results indicate that vowel quality does indeed have a significant effect. For three of the four lexical tones of Mandarin, vowel quality facilitates tone recognition—tones were recognized faster when coproduced with /a/ and /i/ than when coproduced with /u/. Past work has shown that tones exert coarticulatory influences on the tongue body in the production of /a/ and /i/ but not in the production of /u/ (Shaw *et al.*, 2016). The direction of the facilitatory effects of vowel quality on tone recognition is therefore consistent with the hypothesis that tone-conditioned variation in the articulation of /a/ and /i/ facilitates rapid recognition of tones. We speculate that the one exception to this general pattern—no effect of vowel quality on falling tone perception—is due to fortuitous amplification of falling tone harmonics by the segment sequence /tu/. This explanation offers another possible means through which resonances

associated with supralaryngeal constrictions for segments can influence tone perception.

ACKNOWLEDGMENTS

We would like to thank Benjawan Kasisopa and Denis Burnham for sharing the Mandarin recordings that were used as stimuli; Yuan Ma and Chong Han for help recruiting participants; and Donald Derrick, Michael Proctor, Weirong Chen, Denis Burnham, Seth Wiener, as well as audiences at Interspeech Lyon and the University of Potsdam, for comments and helpful discussions at various stages in the development of this research. Any errors and oversights are the sole responsibility of the authors.

[1]Here and throughout we adopt the notional conventions for formants, harmonics, and resonances proposed by Titze *et al*. (2015).
[2]Gating tasks provide another experimental paradigm for looking at the timecourse of tone recognition (Lai and Zhang, 2008) and also show effects of tone probability on tone recognition (Wiener and Ito, 2016).
[3]Adding the main effect of tone without the interaction with condition did not improve the model.
[4]The thickness of the red lines representing the formants is the log of the bandwidth divided by two.

Abramson, A. S. (**1962**). *The Vowels and Tones of Standard Thai: Acoustical Measurements and Experiments* (Indiana University Research Center in Anthropology, Folklore, and Linguistics, Bloomington, IN).
Abramson, A. S. (**1977**). "Noncategorical perception of tone categories in Thai," J. Acoust. Soc. Am. **61**(Suppl. 1), S66.
Barr, D. J. (**2008**). "Analyzing 'visual world' eyetracking data using multilevel logistic regression," J. Mem. Lang. **59**, 457–474.
Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Software **67**, 1–48.
Beddor, P. S., Harnsberger, J. D., and Lindemann, S. (**2002**). "Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates," J. Phonetics **30**, 591–627.
Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., and Brasher, A. (**2013**). "The time course of perception of coarticulation," J. Acoust. Soc. Am. **133**, 2350–2366.
Benguerel, A.-P., and McFadden, T. U. (**1989**). "The effect of coarticulation on the role of transitions in vowel perception," Phonetica **46**, 80–96.
Boersma, P., and Weenink, D. (**2013**). "Praat: Doing phonetics by computer (version 5.3.82) [computer program]," http://www.praat.org (Last viewed 04/15/2020).
Brunelle, M. (**2009**). "Contact-induced change? Register in three Cham dialects," J. Southeast Asian Ling. Soc. **2**, 1–22.
Chen, L.-M., Chao, K.-Y., and Peng, J.-F. (**2007**). "VOT productions of word-initial stops in Mandarin and English: A cross-language study," in *Proceedings of the 19th Conference on Computational Linguistics and Speech Processing*, edited by K.-H. Chen and B. Chen (The Association for Computational Linguistics and Chinese Language Processing, Taipei, Taiwan), pp. 303–317.
Connell, B. (**2002**). "Tone languages and the universality of intrinsic F0: Evidence from Africa," J. Phonetics **30**, 101–129.
Cynx, J., and Shapiro, M. (**1986**). "Perception of missing fundamental by a species of songbird (*Sturnus vulgaris*)," J. Comp. Psychol. **100**, 356–360.
Edmondson, J. A., Ziwo, L., Esling, J. H., Harris, J. G., and Li, S. (**2001**). "The aryepiglottic folds and voice quality in the Yi and Bai languages: Laryngoscopic case studies," Mon Khmer Stud. **31**, 83–110.
Edmondson, J. A., and Gregerson, K. J. (**1993**). "Western Cham as a register language," Ocean. Ling. Spec. Publ. **24**, 61–74.
Edwards, J., and Harris, K. S. (**1990**). "Rotation and translation of the jaw during speech," J. Speech Lang. Hear. Res. **33**, 550–562.
Erickson, D., Iwata, R., Endo, M., and Fujino, A. (**2004**). "Effect of tone height on jaw and tongue articulation in Mandarin Chinese," in *Proceedings of the International Symposium on the Tonal Aspects of Language: With Emphasis on Tone Languages (TAL-2004)*, edited by B. Bel and I. Marlien (Chinese Academy of Sciences, Beijing, China), pp. 53–56.
Gandour, J. (**1977**). "On the interaction between tone and vowel length: Evidence from Thai dialects," Phonetica **34**, 54–65.
Gao, M. (**2009**). "Gestural coordination among vowel, consonant and tone gestures in Mandarin Chinese," Chin. J. Phonetics **2**, 43–50.
Gordon, M. (**2001**). "A typology of contour tone restrictions," Stud. Lang. **25**, 423–462.
Greenberg, S., and Zee, E. (**1979**). "On the perception of contour tones," UCLA Work. Pap. Phonetics **45**, 150–164.
Heffner, H., and Whitfield, I. C. (**1976**). "Perception of the missing fundamental by cats," J. Acoust. Soc. Am. **59**, 915–919.
Hirahara, T., and Kato, H. (**1992**). "The effect of F0 on vowel identification," in *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Ohmsha/IOS, Tokyo), pp. 89–112.
Honda, K. (**1995**). "Laryngeal and extra-laryngeal mechanisms of F0 control," in *Producing Speech: Contemporary Issues for Katherine Safford Harris*, edited by F. Bell-Berti and L. J. Raphael (AIP Press, Woodbury, NY, USA), pp. 215–232.
Hoole, P., and Hu, F. (**2004**). "Tone-vowel interaction in standard Chinese," in *Proceedings of the International Symposium on the Tonal Aspects of Language: With Emphasis on Tone Languages (TAL-2004)*, edited by B. Bel and I. Marlien (Chinese Academy of Sciences, Beijing, China), pp. 89–92.
House, D. (**1996**). "Differential perception of tonal contours through the syllable," in *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 96)*, edited by H. T. Bunnell and W. Idsardi (Alfred. I. DuPont Institute, Philadelphia, PA, USA), pp. 2048–2051.
Howie, J. M. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge University Press, New York).
Hu, J., Gao, S., Ma, W., and Yao, D. (**2012**). "Dissociation of tone and vowel processing in Mandarin idioms," Psychophysiology **49**, 1179–1190.
Huang, Y. (**2019**). "The role of creaky voice attributes in Mandarin tonal perception," in *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, edited by S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Australasian Speech Science and Technology Association Inc., Canberra, Australia), pp. 1465–1469.
Iseli, M., Shue, Y. L., and Alwan, A. (**2007**). "Age, sex, and vowel dependencies of acoustic measures related to the voice source," J. Acoust. Soc. Am. **121**(4), 2283–2295.
Jiang-King, P. (**1999**). *Tone-Vowel Interaction in Optimality Theory* (LINCOM EUROPA, Munich, Germany).
Kreiman, J., and Gerratt, B. R. (**2010**). "Perceptual sensitivity to first harmonic amplitude in the voice source," J. Acoust. Soc. Am. **128**, 2085–2089.
Kreiman, J., Gerratt, B. R., and Kahn, S. D. (**2010**). "Effects of native language on perception of voice quality," J. Phonetics **38**, 588–593.
Kuang, J. (**2017**). "Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice," J. Acoust. Soc. Am **142**, 1693–1706.
Kuang, J., and Liberman, M. (**2018**). "Integrating voice quality cues in the pitch perception of speech and non-speech utterances," Front. Psychol. **9**, 2147.
Lai, Y., and Zhang, J. (**2008**). "Mandarin lexical tone recognition: The gating paradigm," Kansas Work. Pap. Ling. **30**, 183–194.
Lee, C. Y., Tao, L., and Bond, Z. S. (**2008**). "Identification of acoustically modified Mandarin tones by native listeners," J. Phonetics **36**, 537–563.
Lisker, L., and Abramson, A. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**, 384–422.
Liu, S., and Samuel, A. G. (**2007**). "The role of Mandarin lexical tones in lexical access under different contextual conditions," Lang. Cognit. Processes **22**, 566–594.
Liversedge, S. P., Zang, C., Zhang, M., Bai, X., Yan, G., and Drieghe, D. (**2014**). "The effect of visual complexity and word frequency on eye movements during Chinese reading," Visual Cognit. **22**, 441–457.
Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., and Aslin, R. N. (**2007**). "The dynamics of lexical competition during spoken word recognition," Cognit. Sci. **31**, 133–156.
Malins, J. G., and Joanisse, M. F. (**2010**). "The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study," J. Mem. Lang. **62**, 407–420.

Matin, E., Shao, K., and Boff, K. R. (**1993**). "Saccadic overhead: Information-processing time with and without saccades," Attent. Percept. Psychophys. **53**, 372–380.
Meddis, R., and O'Mard, L. (**1997**). "A unitary model of pitch perception," J. Acoust. Soc. Am. **102**, 1811–1820.
Mills, C. B. (**1980**). "Effects of the match between listener expectancies and coarticulatory cues on the perception of speech," J. Exp. Psychol. Hum. Percept. Perform. **6**, 528–535.
Moisik, S. R., Lin, H., and Esling, J. H. (**2014**). "A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS)," J. Int. Phonetic Assoc. **44**(01), 21–58.
Nittrouer, S., and Studdert-Kennedy, M. (**1987**). "The role of coarticulatory effects in the perception of fricatives by children and adults," J. Speech Lang. Hear. Res. **30**, 319–329.
Ohman, S. (**1966**). "Coarticulation in VCV utterances," J. Acoust. Soc. Am. **39**, 151–168.
Ostry, D. J., and Munhall, K. G. (**1994**). "Control of jaw orientation and position in mastication and speech," J. Neurophysiol. **71**, 1528–1545.
Peng, G., Zheng, H.-Y., Gong, T., Yang, R.-X., Kong, J.-P., and Wang, W. S.-Y. (**2010**). "The influence of language experience on categorical perception of pitch contours," J. Phonetics **38**, 616–624.
Poeppel, D., Phillips, C., Yellin, E., Rowley, H. A., Roberts, T. P., and Marantz, A. (**1997**). "Processing of vowels in supratemporal auditory cortex," Neurosci. Lett. **221**, 145–148.
Porretta, V., Tucker, B. V., and Järvikivi, J. (**2016**). "The influence of gradient foreign accentedness and listener experience on word recognition," J. Phonetics **58**, 1–21.
R Development Core Team. (**2006**). "*R*: A language and environment for statistical computing," *R* Foundation for Statistical Computing, Vienna, Austria, available at https://www.r-project.org/ (Last viewed 04/15/2020).
Rose, P. J. (**1988**). "On the non-equivalence of fundamental frequency and pitch in tonal description," in *Prosodic Analysis and Asian Linguistics to Honour R. K. Sprigg. Pacific Linguistics, C-104*, edited by D. Bradley, E. J. A. Henderson, and M. Mazaudon (Dept. of Linguistics, Research School of Pacific and Asian Studies, Australian National University, Canberra, Australia), pp. 55–82.
Schirmer, A., Tang, S.-L., Penney, T. B., Gunter, T. C., and Chen, H.-C. (**2005**). "Brain responses to segmentally and tonally induced semantic violations in Cantonese," J. Cognit. Neurosci. **17**, 1–12.
Shadle, C. H., Nam, H., and Whalen, D. H. (**2016**). "Comparing measurement errors for formants in synthetic and natural vowels," J. Acoust. Soc. Am. **139**, 713–727.
Shaw, J. A., and Chen, W. R. (**2019**). "Spatially-conditioned speech timing: Evidence and implications," Front. Psychol. **10**, 2726.
Shaw, J. A., Chen, W.-R., Proctor, M. I., and Derrick, D. (**2016**). "Influences of tone on vowel articulation in Mandarin Chinese," J. Speech Lang. Hear. Res. **59**, S1566–S1574.
Shen, J., Deutsch, D., and Rayner, K. (**2013**). "On-line perception of Mandarin tones 2 and 3: Evidence from eye movements," J. Acoust. Soc. Am. **133**, 3016–3029.
Shi, B., and Zhang, J. (**1987**). "Vowel intrinsic pitch in Standard Chinese," in *Proceedings of the 11th International Congress of Phonetic Sciences* (Academy of Sciences of the Estonian SSR, Tallinn, Estonia), pp. 142–145.
Shuai, L., and Malins, J. G. (**2017**). "Encoding lexical tones in jTRACE: A simulation of monosyllabic spoken word recognition in Mandarin Chinese," Behav. Res. Methods **49**, 230–241.
Sun, K.-C., and Huang, T. (**2012**). "A cross-linguistic study of Taiwanese tone perception by Taiwanese and English listeners," J. East Asian Ling. **21**, 305–327.
Titze, I. R., Baken, R. J., Bozeman, K., Granqvist, S., Henrich, N., Herbst, C. T., Howard, D. M., Hunter, E. J., Kaelin, D., Kent, R., Kreiman, J., Kob, M., Löfqvist, A., McCoy, S., Miller, D. G., Noe, H., Scherer, R. C., Smith, J. R., Story, B. H., Svec, J. G., Ternström, S., and Wolfe, J. (**2015**). "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," J. Acoust. Soc. Am. **137**, 3005–3007.
Tomlinson, R. W., and Schwarz, D. W. (**1988**). "Perception of the missing fundamental in nonhuman primates," J. Acoust. Soc. Am. **84**, 560–565.
Tong, Y., Francis, A. L., and Gandour, J. T. (**2008**). "Processing dependencies between segmental and suprasegmental features in Mandarin Chinese," Lang. Cognit. Processes **23**, 689–708.
Whalen, D. H., Gick, B., Kumada, M., and Honda, K. (**1999**). "Cricothyroid activity in high and low vowels: Exploring the automaticity of intrinsic F0," J. Phonetics **27**, 125–142.
Whalen, D. H., and Levitt, A. G. (**1995**). "The universality of intrinsic F0 of vowels," J. Phonetics **23**, 349–366.
Wiener, S., and Ito, K. (**2015**). "Do syllable-specific tonal probabilities guide lexical access? Evidence from Mandarin, Shanghai and Cantonese speakers," Lang. Cognit. Neurosci. **30**, 1048–1060.
Wiener, S., and Ito, K. (**2016**). "Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners," J. Phonetics **56**, 38–51.
Wu, J., Chen, Y., van Heuven, V. J., and Schiller, N. O. (**2019**). "Dynamic effect of tonal similarity in bilingual auditory lexical processing," Lang. Cognit. Neurosci. **34**, 580–598.
Wu, Z., and Ortega-Llebaria, M. (**2017**). "Pitch shape modulates the time course of tone vs pitch-accent identification in Mandarin Chinese," J. Acoust. Soc. Am. **141**, 2263–2276.
Xu, Y. (**1994**). "Production and perception of coarticulated tones," J. Acoust. Soc. Am. **95**, 2240–2253.
Xu, Y. (**2009**). "Timing and coordination in tone and intonation: An articulatory-functional perspective," Lingua **119**, 906–927.
Xu, Y., Gandour, J. T., and Francis, A. L. (**2006**). "Effects of language experience and stimulus complexity on the categorical perception of pitch direction," J. Acoust. Soc. Am. **120**, 1063–1074.
Xu Rattanasone, N., Attina, V., Kasisopa, B., and Burnham, D. (**2014**). "How to compare tones," in *South and Southeast Asian Psycholinguistics*, edited by H. Winskel and P. Padakannaya (Cambridge University Press, Cambridge, UK), pp. 233–246.
Ye, Y., and Connine, C. M. (**1999**). "Processing spoken Chinese: The role of tone information," Lang. Cognit. Processes **14**, 609–630.
Yip, M. (**2002**). *Tone* (Cambridge University Press, Cambridge, UK).
Zee, E. (**1980**). "Tone and vowel quality," J. Phonetics **8**, 247–258.
Zhang, J. (**2002**). *The Effects of Duration and Sonority on Contour Tone Distribution: A Typological Survey and Formal Analysis* (Routledge, New York).
Zhang, X., Samuel, A. G., and Liu, S. (**2012**). "The perception and representation of segmental and prosodic Mandarin contrasts in native speakers of Cantonese," J. Mem. Lang. **66**, 438–457.