Toward an Understanding of Heritage Prosody: Acoustic and Perceptual Properties of Tone Produced by Heritage, Native, and Second Language Speakers of Mandarin

Charles B. Chang
Boston University

Yao Yao
The Hong Kong Polytechnic University

ABSTRACT

In previous work examining heritage language phonology, heritage speakers have often patterned differently from native speakers and late-onset second language (L2) learners with respect to overall accent and segmentals. The current study extended this line of inquiry to suprasegmentals, comparing the properties of lexical tones produced by heritage, native, and L2 speakers of Mandarin living in the U.S. We hypothesized that heritage speakers would approximate native norms for Mandarin tones more closely than L2 speakers, yet diverge from these norms in one or more ways. We further hypothesized that, due to their unique linguistic experience, heritage speakers would sound the most ambiguous in terms of demographic background. Acoustic data showed that heritage speakers approximated native-like production more closely than L2 speakers with respect to the pitch contour of Tone 3, durational shortening in connected speech, and rates of Tone 3 reduction in non-phrase-final contexts, while showing the highest levels of tonal variability among all groups. Perceptual data indicated that heritage speakers' tones differed from native and L2 speakers' in terms of both intelligibility and perceived goodness. Consistent with the variability results, heritage speakers were the most difficult group to classify demographically. Taken together, these findings suggest that, with respect to tone, early heritage language experience can, but does not necessarily, result in a phonological advantage over L2 learners. Further, they add support to the view that heritage speakers are language users distinct from both native and L2 speakers.

Keywords: *duration, pitch contour, intelligibility, goodness, sociolinguistic classifiability*

INTRODUCTION

How do heritage speakers, or individuals "raised in a home where one language is spoken who subsequently switch to another dominant language" (Polinsky & Kagan, 2007, p. 368), differ phonologically from native speakers as well as adult second language (L2) learners of a language? Although the literature on such switched-dominance bilinguals evinces a recurring theme of divergences from native-speaker norms attributed to "incomplete acquisition" and/or attrition of the heritage language (HL) due to early onset of the dominant language (e.g., Montrul, 2002, 2004), studies of phonetic and phonological knowledge specifically have documented a wide range of linguistic consequences of HL experience, ranging from native-like performance (e.g., Chang, 2016; Lee-Ellis, 2012; Lukyanchenko & Gor, 2011) to novice-like performance (e.g., Pallier, Dehaene, Poline, LeBihan, Argenti, Dupoux, & Mehler, 2003; Ventureyra, Pallier, & Yoo, 2004) to intermediate performance between native and novice (e.g., Lee-Ellis, 2012).

Much of the observed variation in HL performance is attributable to variability in HL speakers' previous experience with, and proficiency in, the HL (Polinsky & Kagan, 2007; Rao, 2015).[1] For example, in the studies of Pallier, Ventureyra and colleagues (cited above), it is international adoptees, or individuals whose HL experience was fully confined to only the first few years (or even months) of life, who pattern like total novices (cf. Oh, Au, & Jun, 2010), whereas individuals with more extensive (and intermittent) experience hearing and/or speaking the HL tend to show measurable advantages over total beginners in (re)learning the HL (e.g., Au, Knightly, Jun, & Oh, 2002; Knightly, Jun, Oh, & Au, 2003).

However, another source of variation in HL performance has to do with the specific measure of performance: the same population of HL speakers does not necessarily pattern the same relative to native speakers and L2 learners across measures, but instead tends to show differences in patterning when tested over a range of phonological variables. For example, HL "overhearers" of Korean (i.e., individuals with experience hearing, but not speaking, the HL) have been found to show an advantage over L2 learners of Korean in stop perception, but not in stop production (Oh, Jun, Knightly, & Au, 2003), while HL speakers of Russian are observed to show native-like levels of phoneme discrimination for some HL contrasts, but not for others (Lukyanchenko & Gor, 2011). One way in which phonological contrasts are known to differ is timescale: suprasegmental contrasts, such as tone and intonation patterns, unfold over a longer time span than segmental contrasts, which dovetails with neural evidence of different time windows at work in speech perception (Obrig, Rossi, Telkemeyer, & Wartenburger, 2010; Poeppel & Hackl, 2008). Nevertheless, the literature on HL phonology is dominated by studies of the segmental level, leaving the suprasegmental level relatively underexplored.

The study reported in this paper is an attempt to examine the understudied domain of suprasegmental production by HL speakers, with a view toward better understanding how HL speakers may resemble and/or differ from native speakers and L2 learners. Specifically, we investigate the production of lexical tone by HL speakers of Mandarin Chinese in comparison to native Mandarin speakers and adult L2 learners of Mandarin, and use a two-pronged (i.e., acoustic and perceptual) approach to gain broad insight into HL tone production. In the following section, we provide a primer on Mandarin tone, review the brief literature on HL suprasegmentals (with a focus on Mandarin), and motivate four research questions about HL Mandarin speakers' speech production.

BACKGROUND

Tone in Mandarin Chinese

The tonal inventory of Mandarin contains four main tones, plus a fifth, "neutral" tone (i.e., Tone 0, or T0), which is restricted to weak, unstressed syllables and, thus, often analyzed as the absence of one of the four main tones rather than a full-fledged tone itself (Duanmu, 2007). Whereas T0 does not occur in isolation, the four main tones do, and their pitch contours in isolation are standardly taken to be their canonical shapes: a high flat contour for Tone 1 (T1), a mid-to-high rising contour for Tone 2 (T2), a low falling-rising contour for Tone 3 (T3), and a high-to-low falling contour for Tone 4 (T4).

The primary cue to tone in Mandarin is pitch; however, there are also secondary cues such as duration, phonation, and amplitude properties, which allow native perception to remain quite accurate when the acoustic correlate of pitch, fundamental frequency (f_0), is unavailable (Kong & Zeng, 2006; Liu & Samuel, 2004). With regard to duration, the isolation form of T3, for example, is significantly longer than that of the other tones, whereas the isolation form of T4 is significantly shorter; in addition, both T3 and T4 often include intervals of creaky phonation, unlike T1 and T2 (Chang & Yao, 2007; Chao, 1933; Kuang, 2013).

Although each of T1–T4 has a distinct pitch contour, these contours differ in terms of their phonetic similarity to one another. In particular, T2 and T3 are observed to be highly confusable for both native and non-native listeners (Hao, 2012; Kiriloff, 1969; Shen & Lin, 1991), which may be attributed to the fact that the contour for T2 (generally described as a "rising" tone) typically falls before it rises, much like the contour for T3. As a result, both tones may be characterized acoustically as containing a "turning point" (i.e., change in direction) in their f_0 contour; however, they differ in terms of the timing of this turning point, with T2 showing an earlier turning point than T3 (Shi & Wang, 2006). This difference in the timing of the turning point has been found to be an important cue to the perceptual distinction between T2 and T3 (Shen & Lin, 1991; Shen, Lin, & Yan, 1993).

The pitch contours of T1–T4, while all having a canonical shape, also vary considerably across contexts due to coarticulation (Chang & Bowles, 2015; Xu, 1997) as well as alternation. In regard to alternation, T3 in particular occurs in two forms: a "full" (i.e., falling-rising) contour, which occurs before a pause (especially in isolation), and a "half" (i.e., falling only) contour, which occurs before any of the other tones. When occurring before pause while preceded by another tone, T3 may occur as half or full T3; however, half T3 is more common, as full T3 in this context has an emphatic connotation (Duanmu, 2007, pp. 238–239).[2] A number of tone sandhi rules in Mandarin contribute further variability in tone realization by causing one tone to surface as a different tone in certain contexts (e.g., T3 > T2 before another T3; 不 /pu˨˩˦/ 'not' with T4 > [pu˧˥] with T2 before another T4); however, these rules are not of concern here because the specific contexts and/or lexical items to which these rules apply were not included in the current study.

Perception and Production of Tone by Heritage Speakers

Given the asymmetry in the HL phonological literature between segmental and suprasegmental studies, published findings on the perception or production of tone by individuals with HL experience in a tonal language are few, coming mainly from two studies addressing HL speakers' performance on Mandarin tonal contrasts: Yang (2015) and Tsukada, Xu and Xu Rattanasone (2015).

Yang's (2015) study contains, to our knowledge, the only published data on HL speakers' performance on tonal contrasts in their specific HL. This study examined the perception and production of Mandarin tones by native speakers, late-onset L2 learners, and HL speakers (specifically, relearners), arguing that HL speakers' tone perception and production patterns are both intermediate (i.e., between the patterns for native speakers and L2 learners). In perception, this was the case with respect to categoricalness and stability of tone perception; additionally,

HL speakers resembled L2 learners in showing greater overall reliance on register (i.e., pitch level) than native speakers, while at the same time resembling native speakers in their ability to recognize the starting pitch level of a tone. In production, intermediate patterning was found in the overall production space as well as in pitch range, which was larger for HL speakers than for L2 learners. These results provide a solid starting point for understanding HL tone production; however, because the profile of HL experience for the HL group in this study was not described, the generalizability of the results is unclear.

The study of Tsukada, Xu and Xu Rattanasone (2015) focuses on perception and reports that, in terms of tone discrimination, HL speakers are either similar to, or less accurate than, late-onset L2 learners. However, the HL experience of these HL speakers was not actually with the target language (i.e., Mandarin), but a different variety of Chinese with a different tone system (i.e., Cantonese, which contains 6–9 tones, depending on the analysis). Thus, this HL group might be better regarded as third language (L3) learners rather than HL speakers in the context of the current study.

Research Questions and Predictions

Given that Yang (2015) provides the only published data on HL speakers' perception and/or production of tone in their specific HL, further investigation of HL speakers' tone production is needed, especially because it is not clear how Yang's results (obtained with an underspecified sample of HL Mandarin speakers) may generalize to HL Mandarin speakers at large, a group known to be highly heterogeneous in terms of HL proficiency (Li & Duff, 2008). Consequently, we conducted an extensive acoustic and perceptual investigation of HL Mandarin speakers' tone production with a speaker sample evincing a wide range of HL experience in order to address four research questions.

The first question is how HL speakers compare to native speakers and adult L2 learners with respect to acoustic properties of their tone production. The specific acoustic properties examined include the duration and fundamental frequency (f_0) contour of the tones, the f_0 range observed over all tones, and the turning point of tones that change in pitch direction (T2, T3). Given prior findings on segmental production in the HL as well as the results of Yang (2015), we hypothesize that HL speakers will pattern closely with native speakers for some properties, but closely with L2 learners for other properties; that is, we expect the patterning of HL speakers relative to native speakers and L2 learners to differ across acoustic properties. As for the specific patterning for each acoustic property, the literature supports only one prediction: on the basis of Yang's results, we predict that HL speakers will, in contrast to L2 learners, pattern closely with native speakers in terms of f_0 range.

The second question is how HL speakers measure up to native speakers and adult L2 learners in terms of tonal variability. Because of the heterogeneity of HL speakers' experience with Mandarin, which often includes little explicit instruction in tonal targets (in contrast to the typical experience of educated native speakers and instructed L2 learners), we predict that HL speakers will have more diffuse articulatory targets for isolated tones and, consequently, show more variability in their citation-form tone contours than native or L2 speakers.

The third question is how HL speakers' tones are perceived by native Mandarin listeners. Given the lack of clear predictions regarding the patterning of HL speakers with respect to specific acoustic properties and the prediction of higher tonal variability for HL speakers, we predict that HL speakers' tones will, on average, be more difficult to identify (i.e., less intelligible) than native speakers' tones, but not more difficult to identify than L2 speakers' tones (which are likely to be even more divergent from native norms). At the same time, our prediction of greater tonal variability for HL speakers also leads us to expect that when HL speakers' tones are identified correctly, this will often be due to a relatively native-like tone contour (as opposed to a consistently produced non-native-like contour); consequently, we predict that when HL speakers' tones are intelligible, they will rate higher in terms of goodness (i.e., native-likeness) compared to the intelligible tones of L2 learners.

The fourth question is how HL speakers are perceived sociolinguistically by native Mandarin listeners. In particular, we are interested in whether, on the basis of their speech alone, native listeners will find it relatively difficult to identify HL Mandarin speakers as such, compared to other native speakers or to L2 learners. Our prediction of greater tonal variability for HL speakers leads us to expect that HL speakers' speech will be more ambiguous (in terms of the demographic background of the talker) than native or L2 speakers' speech. Thus, we predict that, from among native, L2, and HL speakers, the most difficult group for native listeners to classify sociolinguistically (i.e., demographically) will be HL speakers.

METHODOLOGY

Participants

The 26 Mandarin talkers who participated in the production experiment consisted of three groups differing in terms of prior experience with the language (the same speakers examined in Chang, Yao, Haynes, & Rhodes, 2011): a group of native Mandarin (NM) speakers, a group of late-onset L2 learners, and a group of HL speakers.

The NM group (n = 6, 4 females, 2 males, mean age 29.8 years, SD = 8.5) comprised NM speakers who were long-term residents of the U.S. with communicative competence in English. Note that we use the term *native* here to refer to the fact that these speakers acquired Mandarin from birth through adolescence; in particular, we do not use it to mean "monolingual" since the appropriate native comparison group in this case consists of non-monolinguals. This is because the Mandarin input to which the HL speakers were exposed (i.e., the ostensible target variety) came primarily from Mandarin-speaking relatives in the U.S., who were likely to be familiar with English (and, therefore, were not monolingual). NM participants were all born and educated in a Mandarin-speaking region (namely, Mainland China or Taiwan) up until at least seventh grade, reported their current Mandarin proficiency level to be native-like, and judged Mandarin to be their best language. At the time of study, all were either students or visiting scholars at the University of California, Berkeley, with a mean age of arrival (AoA) to the U.S. of 24.2 years (SD = 8.1); consequently, all spoke English in addition to Mandarin, with two reporting knowledge of another variety of Chinese (Cantonese, Shanghainese) as well.

The L2 group (n = 5, 3 females, 2 males, mean age 21.6 years, SD = 3.7) comprised adult learners of Mandarin who had acquired the language through formal instruction and/or prior

travel to a Mandarin-speaking country. L2 participants were native speakers of American English who were born and educated in the U.S., were raised in English-speaking families (monolingual English for three, English plus another non-tonal language for two), and started to learn Mandarin after the age of 18. The amount of prior Mandarin experience ranged from 2.5 weeks of immersion to 2 years of foreign language instruction, which constituted these speakers' only experience with any tonal language.[3] L2 participants generally described their Mandarin proficiency at the time of the experiment as relatively poor, with self-reported estimates of conversational comprehension ranging from 10% to 50%.

The HL group comprised Chinese Americans who were born to Mandarin-speaking parents and thus had some degree of prior Mandarin experience in the home, but who reported speaking English most of the time overall and did not fulfill all of the criteria for inclusion in the NM group (i.e., being raised continuously in a Mandarin-speaking country until adolescence, perceiving their Mandarin proficiency to be native-like, and identifying Mandarin as their dominant language). Because of the wide range observed in their previous Mandarin exposure, HL participants were further assigned to one of two subgroups based on frequency of current Mandarin use and amount of time spent in a Mandarin-speaking country: a high-exposure (HE) subgroup (n = 9, 4 females, 5 males, mean age 21.0 years, SD = 1.7) and a low-exposure (LE) subgroup (n = 6, 4 females, 2 males, mean age 20.0 years, SD = 1.1). HE participants reported using Mandarin to communicate with both parents most or all of the time, with most (7/9) having been born and/or resided in a Mandarin-speaking country for a significant portion of their childhood (mean AoA to U.S. = 6.9 years). By contrast, LE participants reported using Mandarin at home half of the time or less and, with one exception, had never lived in a Mandarin-speaking country. See Chang, Yao, Haynes and Rhodes (2011) for further details on residential history, language exposure, and HL use of the participants in this group.

The Mandarin listeners who served as judges in the perceptual rating experiment comprised 64 NM speakers (47 females, 17 males, mean age 23.7 years, SD = 4.2) who were born, raised, and educated primarily in Mainland China. Representing diverse regions of origin within Mainland China, almost all were also familiar with a non-standard Chinese dialect (as is the case for most Mandarin speakers); however, the distribution of regions of origin in the dataset did not allow for an analysis of dialectal background in relation to ratings. At the time of the study, the listeners were pursuing a degree program at a university in Hong Kong. None had prior experience with teaching Chinese to L2 learners and none reported a history of speech or hearing disorders.

Materials

The materials for the Mandarin production task consisted of a total of 59 items, of which 22 were critical items and 37 were fillers and items included as part of other studies not discussed here (including that reported in Chang, Yao, Haynes and Rhodes, 2011). The 22 critical items comprised 16 monosyllabic items in the form of four distinct minimal quadruplets (all containing a postalveolar sibilant onset consonant and a low vowel nucleus), as well as six multisyllabic items: two disyllabic, three trisyllabic, and one quadrisyllabic. The multisyllabic items were constructed to contain common words likely to be familiar to the participants and included T0 in both final and non-final positions and in positions preceding and following each of T1–T4; however, our focus in this study is on the full tones (i.e., T1–T4). The complete set of critical

items is shown in Table 1, where tones are transcribed with Chao tone letters and syllables with the neutral tone are underlined. The abbreviation ASP glosses the aspect marker /lə/.

Table 1.

Critical items used in the production task.

Monosyllabic Items				Multisyllabic Items	
T1	T2	T3	T4		
沙 /ʂa˥/ ‘sand’	啥 /ʂa˧/ ‘what’	傻 /ʂa˨˩/ ‘stupid’	煞 /ʂa˥˩/ ‘suddenly’	桌子 /t͡ʂwo˨˩ ts̬ɨ̞/ ‘table’	儿子 /əɚ˧ ts̬ɨ̞/ ‘son’
扎 /t͡ʂa˥/ ‘prick’	闸 /t͡ʂa˧/ ‘gate’	眨 /t͡ʂa˨˩/ ‘blink’	炸 /t͡ʂa˥˩/ ‘fry’	喝了水 /xə˨˩ lə̲ ʂweɪ̯˨˩/ ‘drink water + ASP’	吃了饭 /t͡ʂʰi˨˩ lə̲ fan˥˩/ ‘eat food + ASP’
插 /t͡ʂʰa˥/ ‘insert’	茶 /t͡ʂʰa˧/ ‘tea’	衩 /t͡ʂʰa˨˩/ ‘underpants’	岔 /t͡ʂʰa˥˩/ ‘bifurcation’	你的书 /ni˨˩ tə̲ ʂu˥/ ‘your book’	
家 /tɕ͡ja˥/ ‘home’	夹 /tɕ͡ja˧/ ‘clip’	假 /tɕ͡ja˨˩/ ‘false’	嫁 /tɕ͡ja˥˩/ ‘marry’	好看的人 /xaʊ̯˨˩ kʰan˥˩ tə̲ ɻən˧/ ‘good-looking person’	

The stimuli for the perceptual rating task consisted of the speech samples recorded in the Mandarin production task. As four tokens of each item were collected, the set of critical stimuli evaluated in this task comprised 2,288 (22 items × 4 tokens × 26 talkers) sound files in all.

Procedure

This study consisted of two main parts: a Mandarin production experiment and a perceptual rating experiment with native listener judges. The production experiment was carried out in the U.S. (California), while the rating experiment took place in Hong Kong.

Talkers in the production study first completed a detailed background questionnaire (adapted from Dai & Zhang, 2008) and then a reading task with an experimenter in a sound-attenuated booth. The questionnaire asked talkers about their residential history and family background, language background, current language use, formal language education, and Mandarin proficiency. In the reading task, talkers were recorded reading the Mandarin items aloud; these items were presented by the experimenter individually on flashcards in random order. Talkers were told to read the items naturally. Each flashcard included an orthographic representation (i.e., Chinese characters) and a romanization (in Pinyin, the system used in Mainland China, and/or Zhuyin/Bopomofo, the system used in Taiwan). The set of 22 critical items was iterated four times, for a total of 88 critical tokens collected for each talker. As described in Chang, Yao, Haynes and Rhodes (2011), the Mandarin production task was part of a larger study that also included an English production task; however, talkers completed all blocks of Mandarin production consecutively, and the order of the Mandarin and English tasks was balanced across talkers. Recordings were made at 48 kHz with 16-bit resolution, using an AKG head-mounted condenser microphone connected either to a Marantz PMD660 recorder or to a Dell desktop computer (through an M-AUDIO USB preamp).

Listeners in the rating study completed one of two types of perception experiments: rating of monosyllables and rating of phrases (both of which consisted of multiple blocks). Each

experiment had four versions, but listeners completed only one version; given the number of listener participants, this resulted in each stimulus being evaluated by approximately eight different listeners. The perception experiments were scripted in Praat (Boersma & Weenink, 2015) and administered with headphones on a Lenovo ThinkPad X240 laptop.

The monosyllable rating experiment consisted of two main blocks. In the first block, listeners made a four-alternative forced-choice identification judgment (where the response options were T1–T4) on the tone in a monosyllabic stimulus and then rated the goodness of that tone on a 1–5 scale. In the second block, listeners heard the same stimuli again and tried to classify the talker's demographic background as one of three options (*native Chinese*, or born and raised in a Chinese-speaking country; *American-born* Chinese,[4] *non-Chinese American*, or not exposed to Chinese until adulthood), and then rated how confident they were of their judgment on a 1–5 scale. To keep the experiment a reasonable length, the full set of stimuli to be rated was divided into four versions of the experiment, such that each version contained one token of every monosyllabic item uttered by every talker.

The phrase rating experiment was overall similar in design to the monosyllable rating experiment, but consisted of four main blocks and differed in terms of the identification response options (which were five: T0–T4). The stimuli were organized into blocks according to syllable count, such that blocks 1–3 focused on disyllabic, trisyllabic, and quadrisyllabic stimuli, respectively. In each of these three blocks, listeners made a forced-choice identification judgment on one of the tones in the given stimulus (played in its entirety, although only one tone was being evaluated), and rated the goodness of that tone. Tones in the first syllable of the stimuli were evaluated first, and then the stimuli of the current block were iterated again so that tones in the next syllable could be evaluated, until the final tones of the stimuli had been reached. In the final block, listeners completed the same demographic classification task as in the final block of the monosyllable rating experiment on all of the multisyllabic stimuli (randomly ordered). As with the monosyllabic stimuli, the full set of multisyllabic stimuli to be rated was distributed among four versions of the experiment.

Data Analysis

Recordings from the production study underwent three stages of acoustic analysis in Praat. In the first stage, the recordings were annotated (by the first author, a trained phonetician) for the onset and offset of the voiced interval over which an f_0 contour would be extracted; this was done via auditory inspection and joint visual inspection of the waveform and a wide-band spectrogram (on the basis of criteria such as changes in periodicity, amplitude, and formant structure), according to the segments in the item. When a target onset consonant was phonologically voiceless (e.g., /ʂ, ʈ/), the onset of the voiced interval was identified with the onset of the following vocoid; in the interest of consistency, this annotation protocol was used regardless of whether the consonant surfaced as voiceless or voiced (e.g., /ʈ/ being produced as [ʈ], [ɖ], or [ð]). When the onset consonant was phonologically voiced (e.g., /ɳ, l/), the onset of the voiced interval was identified with the onset of this voiced consonant. The offset of the voiced interval was identified with the last point of regular visible glottal pulses. Auditory inspection of the recordings at this stage revealed that a small number of tokens (2%) were unsuitable for analysis for one or more reasons (e.g., production errors, false starts, file corruption); these were excluded from further analysis.

In the second stage of the acoustic analysis, measurements of voiced interval durations and of f_0 at 10 evenly spaced time points during each interval (ranging from the 5% point to the 95% point) were extracted via Praat's cross-correlation method. The default settings for this method were used except that the voicing threshold was set to 0.25 and the pitch floor was adjusted according to the talker to provide the best f_0 tracking possible (generally, this was 45 Hz for males and 65 Hz for females).

In the third stage of acoustic analysis, all measurements were manually inspected for f_0 tracking errors. Obvious errors (which occurred in approximately 23% of tokens) were hand-corrected in one of the following two ways. First, the cross-correlation settings were adjusted to correct contours that contained large pitch jumps and/or gaps, and f_0 measurements were taken at the appropriate time point(s) in the corrected contour. However, when the contour resisted correction via adjustment of the analysis settings (usually the case with particularly creaky phonation), an f_0 measurement was calculated manually by taking the duration over a 2–3 period interval around the relevant time point and converting to an f_0 value. All f_0 measurements were then log-transformed and converted to a *T* measure using the formula in (1) (Shi, 1986; Zhu, 2004), where $f_{0\max}$ and $f_{0\min}$ represent, respectively, the highest and lowest f_0 measurements from the talker's production of monosyllabic items (ranging over T1–T4). Thus, the *T* measure of all monosyllabic items was between 0 and 5, comparable to Chao's (1930) five-point tonal representation system, while the *T* measure of multisyllabic items could go outside this range due to the overlay of phrasal intonation.

(1)

$$T_x = \frac{5 \times (\log(f_{0x}) - \log(f_{0\min}))}{\log(f_{0\max}) - \log(f_{0\min})}$$

The statistical analysis of both the acoustic data and the perception data was done with mixed-effects modeling (i.e., linear mixed-effects models for continuous measures and generalized mixed-effects models for ratio/likelihood measures) in R (R Development Core Team, 2015), using the lme4 package (Bates, Maechler, & Bolker, 2011). Models were built first for basic acoustic measures in monosyllabic items: log-transformed voiced interval duration, *T* values at the first and last time points (BeginT and EndT, respectively), average *T* values over all time points (MeanT), and the range of *T* values (RangeT). Fixed-effect predictors were Sex (sum contrast coding) and Group, while random-effect terms were Talker and Item. The critical fixed-effect predictor was Group. In addition to basic acoustic measures, we analyzed a few other measures—including the *T* turning points for T2 and T3, rates of T3 reduction, and other indices of tonal variability—which are described (along with their analyses) in the next section.

Although the production study included four groups (i.e., NM, HE, LE, L2), it is difficult to test the significance of multiple between-group differences (e.g., NM versus HE, HE versus LE, LE versus L2) in a regression model, as the effects are sensitive to the choice of reference level (Clopper, 2013). In view of this, we initially coded the Group variable with three levels: NM, L2, and HL (all HL speakers), where HL subsumed HE and LE and was always set as the reference

level in the model. Thus, the initial models directly tested the difference between HL and NM and the difference between HL and L2, both of which are central to our research questions. If an initial model did not show any significant effect of Group, we rebuilt the model by recoding Group as a binary variable contrasting NM + HE and LE + L2 (reference level). The rationale for this recoding was to test whether there was a split between HE and LE talkers (i.e., HE patterning with NM, LE patterning with L2) that may have led to overall null effects when HL was compared to either NM or L2. To facilitate model interpretation, separate models were built for each tone.

For the perception models, dependent measures were likelihood of accurate tone identification (i.e., intelligibility), goodness rating, (log) combined response time for identification and goodness rating, likelihood of accurate demographic classification (i.e., classifiability), classification confidence rating, and (log) combined response time for classification and confidence rating. Fixed-effect predictors included Group, Tone (sum contrast coding), and their interaction (as above), while random-effect terms were Listener, Talker, and Item. To avoid terms of higher-order interaction (which are difficult to interpret), in all cases, separate models were built for monosyllabic and multisyllabic items. In models of multisyllabic items, additional fixed-effect predictors—length of the phrase in number of syllables (PhraseLen) and sequential position of the current syllable (SyllPos)—were added as control factors. Similar to the acoustic models, perception models were built with alternative ways of coding Group in order to examine both the HL versus NM and HL versus L2 contrasts, as well as the NM + HE versus LE + L2 contrast.

Significance of the predictor terms was determined by *z* values and *p* values in generalized mixed-effects models and by p_{MCMC} values in linear mixed-effects models, with p_{MCMC} values calculated based on the posterior distribution of model parameters generated by the Markov Chain Monte Carlo (MCMC) sampling procedure (10,000 samples; see Baayen, Davidson and Bates, 2008 for a description of the procedure). Predictor terms with *p* or p_{MCMC} values less than .01 were considered to be statistically significant, *p* or p_{MCMC} values between .01 and .05 marginally significant, and *p* or p_{MCMC} values greater than .05 non-significant.

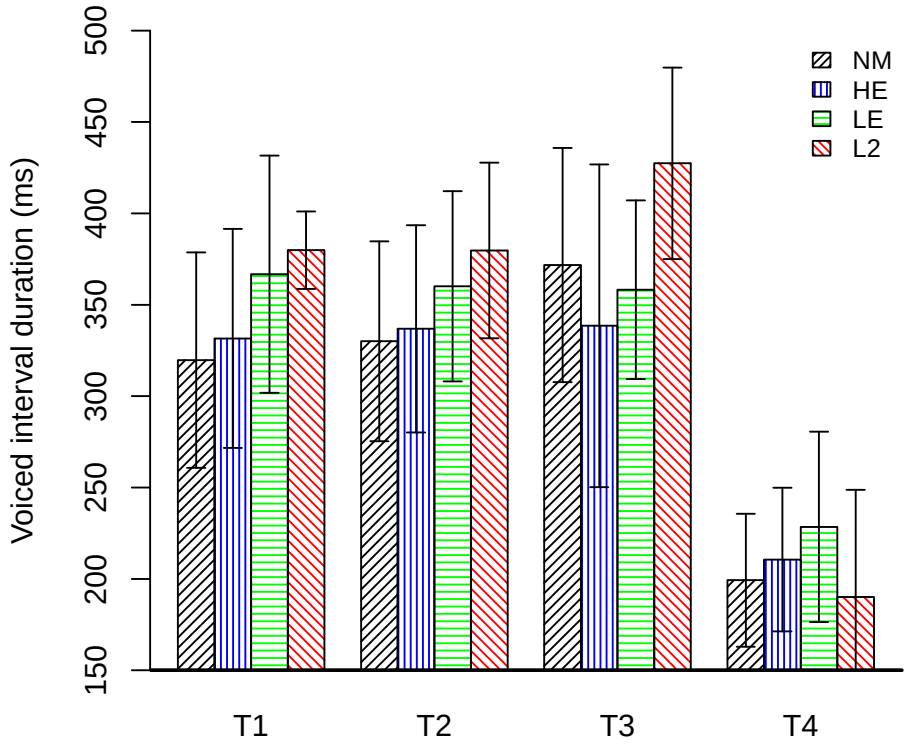
RESULTS

Acoustic Properties: Duration, Pitch Contour, and Variability

We started with analyzing the basic acoustic measures (i.e., voiced interval duration, BeginT, EndT, MeanT, RangeT) in all four tones in monosyllabic items, following the modeling procedure described above. Overall, only three (marginally) significant Group effects were observed across all the models. First, NM + HE talkers produced shorter T1 durations than LE + L2 talkers (β_{NM+HE} = -0.13, t = -2.01, p_{MCMC} = .001), as shown in Figure 1. Second, NM talkers had a slightly lower EndT for T1 than HL talkers (β_{NM} = -0.35, t = -1.93, p_{MCMC} = .003), but a follow-up analysis revealed that the effect was mostly due to one NM talker, who tended to produce a fall toward the end as part of a general vocal pattern of phrase-final glottalization (see Figure 2a); when this talker's data were excluded, the effect of Group on EndT disappeared (β_{NM} = -0.07, t = -0.53, p_{MCMC} = .036). Third, L2 talkers' T3 durations tended to be longer than HL talkers' (β = 0.24, t = 2.15, p_{MCMC} = .002). No other Group effect, with either way of coding

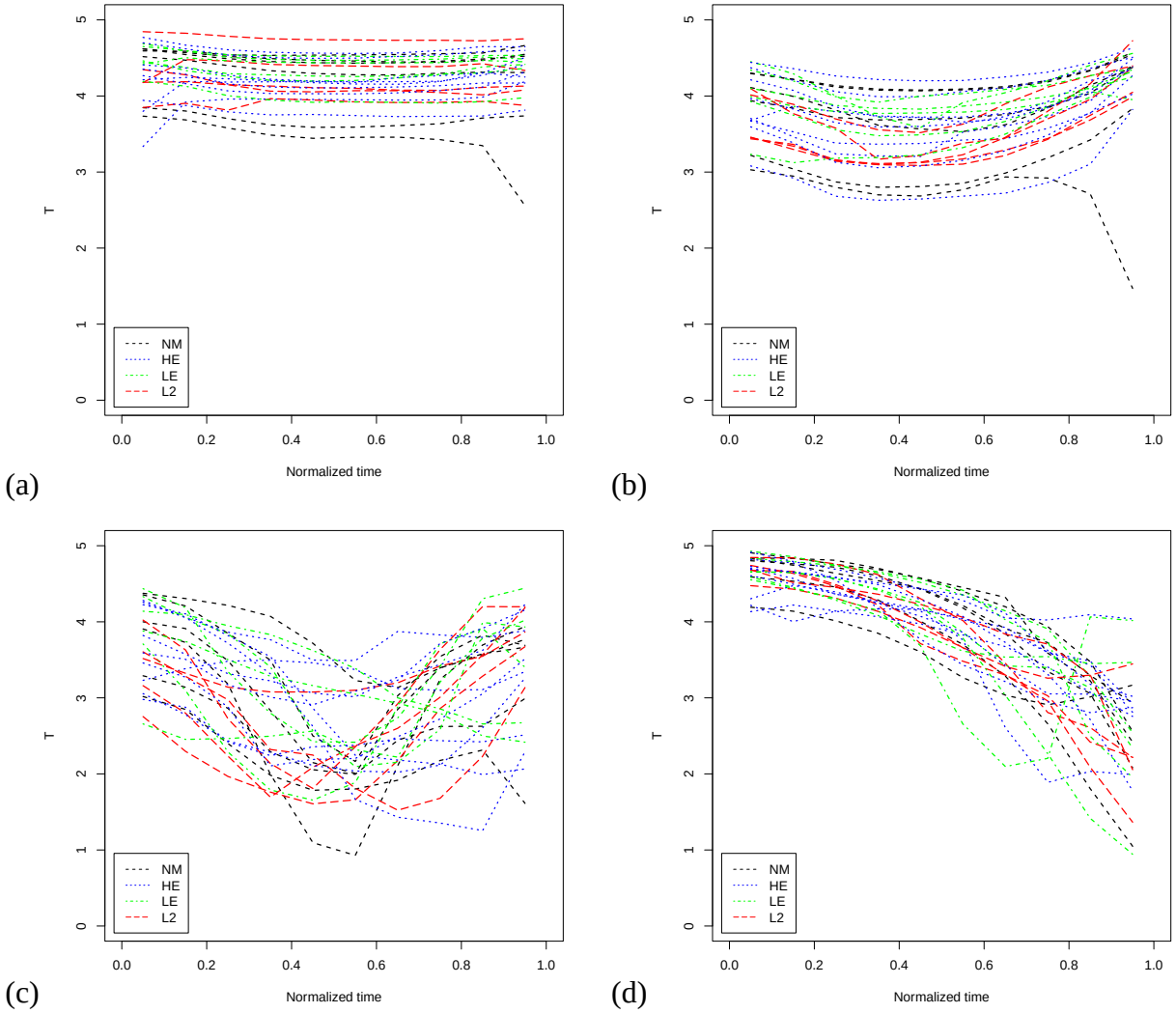
Group described above, was observed in these basic acoustic measures, including RangeT (contra Yang, 2015).

Figure 1. Mean durations (in milliseconds, ms) of T1–T4 in monosyllabic items by talker group (averaged over all talkers' mean values).



Note. Error bars indicate standard error of by-talker mean values.

Figure 2. Mean f_0 contours (in terms of T) by talker (group affiliation indicated by line type and color) for (a) Tone 1, (b) Tone 2, (c) Tone 3, and (d) Tone 4 in monosyllabic items.



The turning point in T2/T3 was identified as the time point (among the 10 time points where *T* was measured) corresponding to the lowest local-minimum *T* value in the tonal contour. Specifically, since the transition from falling to rising contour is usually smooth, we adopted a loose definition of local-minimum: a time point *n* where the corresponding *T* is lower than the *T* value at the previous time point, and lower than or equal to the *T* value at the next time point ($T_{n-1} > T_n \leq T_{n+1}$). All T2 tokens (n = 400) and almost all T3 tokens (397/402) were found to have a valid turning point. Table 2 lists the mean turning points in T2 and T3 by talker group. All groups showed an earlier mean turning point in T2 than in T3 (as expected), but the distance between the turning points of T2 and T3 varied across groups, with the L2 group in particular showing a much smaller difference (0.4) compared to the other three groups (1.1–1.3).

Table 2.

Mean turning points (and standard deviations) in T2 and T3 by talker group. Turning points are measured in terms of time point (1–10, where '1' = 5% and '10' = 95% of the voiced interval).

T2				T3			
NM	HE	LE	L2	NM	HE	LE	L2
4.6 (1.1)	4.5 (1.1)	4.4 (1.0)	4.4 (1.0)	5.7 (1.1)	5.7 (1.8)	5.7 (2.0)	4.8 (1.5)

Mixed-effects models showed significant Group effects on the timing of the T3 turning point, but not of the T2 turning point ($|t| < 1$ with either way of coding Group). For all talker groups, the average turning point in T2 was in between the 4th and 5th time points, corresponding to approximately 40% into its duration.[5] On the other hand, L2 talkers' T3 turned from falling to rising almost one full time point earlier than HL talkers' ($M_{L2} = 4.8$, $M_{HL} = 5.7$; $\beta_{L2} = -0.98$, $t = -2.06$, $p_{MCMC} = .02$), while no overall difference was found between NM and HL talkers ($\beta_{NM} = -0.068$, $t = 0.15$, $p_{MCMC} = .86$). Nevertheless, a post-hoc analysis of within-group variability revealed that HL talkers (both HE and LE) had more variable turning points than NM talkers in T3, as shown in a higher degree of variation both across tokens ($SD_{NM} = 1.1$, $SD_{HE} = 1.8$, $SD_{LE} = 2.0$, $SD_{L2} = 1.5$) and across talkers. All six NM talkers' mean T3 turning points were between 5.2 and 6.2; by contrast, the majority of HE (7/9) and LE (4/6) talkers' means were outside of this range, falling as early as 4.1 and as late as 7.3. On the other hand, L2 talkers—similar to NM talkers—exemplified relatively less variable T3 turning points (all means between 4.3 and 5.8).

Thus, analyses of tonal turning points revealed both that the L2 group had earlier T3 turning points than the NM and HL groups, and that the HL group was especially variable with respect to the T3 turning point. That L2 talkers had earlier T3 turning points may be explained by the fact that L2 learners in classroom contexts receive explicit instruction about the dipping contour of T3 and, therefore, may be particularly eager to reach the pitch trough (and thus, the turning point) of T3's contour and produce a full fall-rise; this is also consistent with the basic acoustic analyses discussed above, which showed that L2 talkers produced T3 with longer durations. HL talkers' high variability with respect to the T3 turning point suggests a possible multimodal distribution, with some initiating T3's rise early in the contour (like L2 talkers), others initiating the rise more in the middle of the contour (like NM talkers), and yet others showing very late turning points, leading to a relatively flat (instead of rising) contour in the remainder of the tone. In other words, the last type of HL talker would be effectively producing half T3 instead of full T3 even when reading a monosyllabic item in isolation.

The high variability observed in HL talkers' T3 turning points is consistent with our second hypothesis, which stated that HL speakers may exhibit higher tonal variability because of more diffuse tonal targets compared to native speakers and L2 learners. To further investigate the issue of tonal variability, we calculated the standard deviation of *T* at each time point across all tokens of the same tone produced by the same talker, and then summed these figures to get an aggregated variability measure (σ) per talker per tone. A series of Welch-corrected two-sample *t*-tests showed that there was only a weak tendency for HL talkers to show greater variability than NM talkers ($t(57.1) = 1.80$, $p = .07$; $M_{HL} = 3.72$, $M_{NM} = 2.83$); furthermore, when T3 was

excluded, the difference between HL and NM talkers became more reliable ($t(53.4) = 2.19$, $p = .03$; $M_{HL} = 2.68$, $M_{NM} = 1.89$). These results thus suggest that the overall difference in variability between HL and NM talkers was *not* driven by HL talkers' high variability on T3; on the contrary, HL talkers had clearly higher tonal variability for the other three tones than NM talkers, and the gap closed on T3 (as both groups showed higher variability on T3 than on the other tones). No other comparisons of σ between groups (HL versus L2, HE versus LE) revealed significant differences.

Due to the small number of multisyllabic items and their unbalanced tonal distribution, we did not examine all the basic acoustic measures for multisyllabic items, but instead focused our analysis on durational shortening (compared with monosyllabic items) and the reduction of T3 in multisyllabic items.[6] As shown in Table 3, tones were overall longer in monosyllabic than multisyllabic items for all groups (all $p < .001$ in two-sample *t*-tests). Among the four tones, T3, as expected, showed the most shortening from monosyllabic to multisyllabic items (mean difference = 66 ms), as the shortening of T3 is part of the general reduction of T3 in many connected speech contexts (we return to this point below). T4 showed the least shortening, probably due to a floor effect, as T4 is the shortest tone in both isolated and connected speech. Consequently, the duration of T1 and T2 emerged as the clearest index of changes in speech rate. A mixed-effects model was built on the (log) durations of all T1 and T2 tokens with Sex and the Group × Context (i.e., monosyllabic, multisyllabic) interaction as fixed-effect predictors, and Talker and Item as random-effect predictors. The model showed no Group effect on durations in monosyllabic tokens ($|t| < 1.3$, $p_{MCMC} > .1$), but a significant Context effect that varied across groups: HL talkers showed a significant effect in the direction of shortening in multisyllabic contexts ($\beta_{multi} = -0.55$, $t = -3.47$, $p_{MCMC} < .001$), which was more pronounced in NM talkers ($\beta_{NM:multi} = -0.06$, $t = -2.48$, $p_{MCMC} = .02$), but less pronounced in L2 talkers ($\beta_{L2:multi} = 0.17$, $t = 7.27$, $p_{MCMC} < .001$). That is, while all groups produced T1 and T2 at similarly slow rates in isolation contexts, NM talkers sped up the most in connected speech, followed by HL talkers and then L2 talkers (in that order).

Table 3.

Mean durations (in ms) of T1–T4 in monosyllabic (mono) versus multisyllabic (multi) items by talker group.

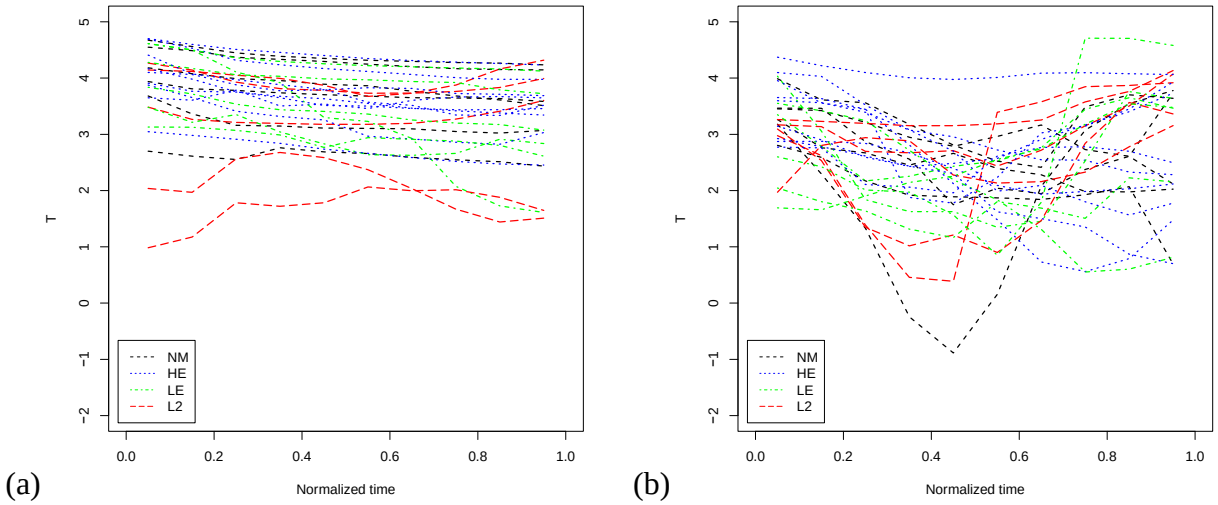
	T1		T2		T3		T4	
	mono	multi	mono	multi	mono	multi	mono	multi
NM	320	160	330	286	372	177	199	164
HE	331	180	337	286	339	187	211	213
LE	367	195	360	279	358	200	228	205
L2	380	246	380	339	427	269	190	224
Grand mean	350	195	352	298	374	208	207	202

To explore the production of T3 in more detail, we examined the degree to which T3 was reduced to half T3 in both biasing and non-biasing contexts by identifying the percentage of tokens that lacked a true turning point. Recall that T3 is consistently realized as half T3 before

any tone except T3, but often (or usually) as full T3 at the end of a prosodic phrase. Two of our multisyllabic stimuli had T3 in initial position followed by T4 and T0, respectively (i.e., non-phrase-final position), while another stimulus item had T3 in phrase-final position. Turning point tracking was attempted for all of these T3 tokens. Because in this case we were specifically interested in the contrast between non-rising (half) and rising (full) realizations of T3, for this analysis, we adopted a more restrictive definition of the turning point (*n*) in the tone contour: $T_{n-1} > T_n < T_{n+1}$.

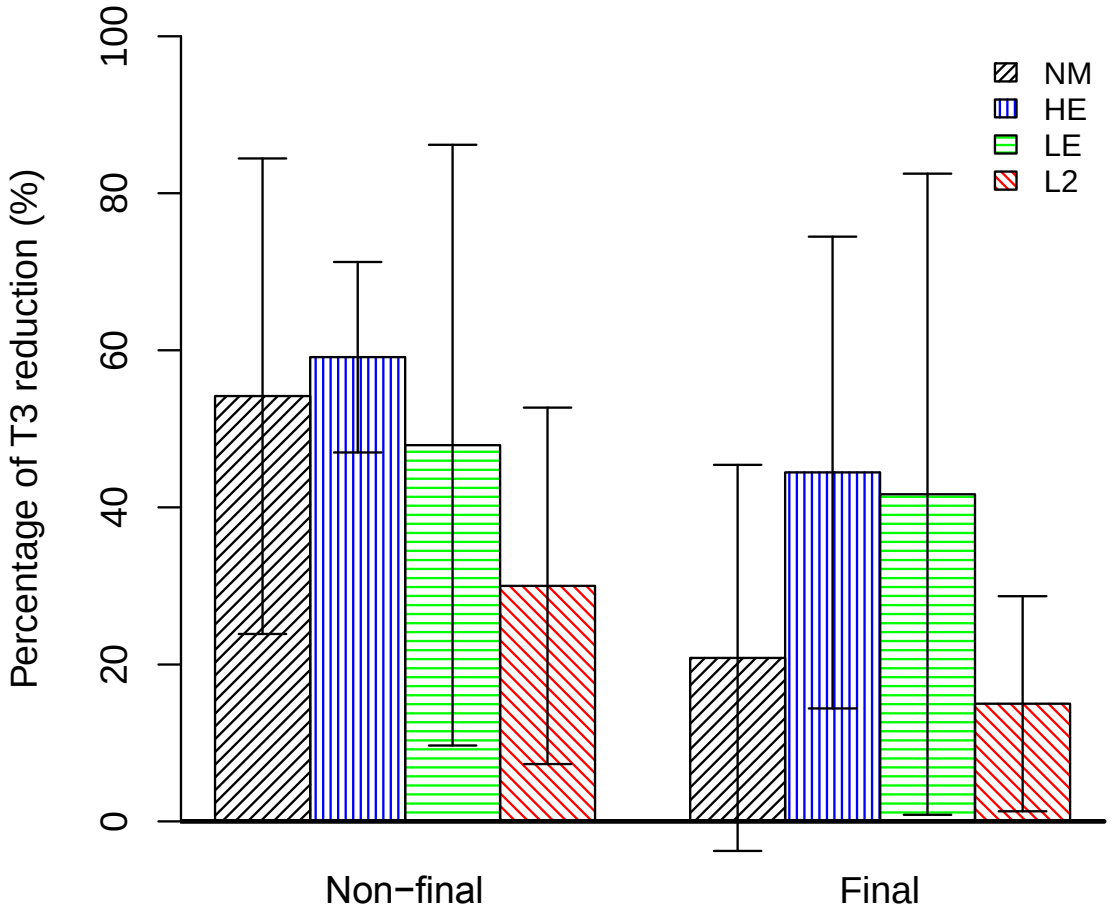
Overall, as expected, there was a higher rate of T3 reduction in non-phrase-final contexts (50%; 103/207) than in phrase-final contexts (33%; 34/103), although both rates were higher than the T3 reduction rate in monosyllabic items (1–15%, depending on the definition of *turning point*). As shown in Figure 3, the contour of T3 differed greatly between the two types of contexts: whereas T3 was often produced with a true turning point in final position (Figure 3b), this was not the case in non-final position, where most talkers produced only a shallow fall (Figure 3a).

Figure 3. Mean f_0 contours (in terms of *T*) by talker (group affiliations indicated by line type and color) for T3 in connected speech: (a) non-phrase-final position, (b) phrase-final position.



Rates of T3 reduction in multisyllabic items also evinced an effect of Group, whereby more Mandarin experience correlated with more T3 reduction (i.e., production of half T3). As shown in Figure 4, in non-phrase-final contexts, the L2 group showed the lowest rate of half T3 production (30%), the LE group the next highest rate (48%), and the HE group the highest rate (HE: 59%; cf. NM: 54%). In phrase-final contexts, the same pattern held for the L2 (15%), LE (42%), and HE (46%) groups, although NM talkers showed an apparently lower T3 reduction rate (21%) than HL talkers. However, when tested by generalized mixed-effects models (fixed effects: Sex, Group; random effects: Talker, Item), only the L2-HL difference was significant, in both non-phrase-final and phrase-final contexts (non-final: β_{L2} = -1.13, z = -2.03, p = .04; final: β_{L2} = -1.57, z = -1.96, p = .05). No reliable difference was found between the NM and HL groups.

Figure 4. Rates of T3 reduction in multisyllabic items, by talker group and context type (averaged over all talkers' mean values).



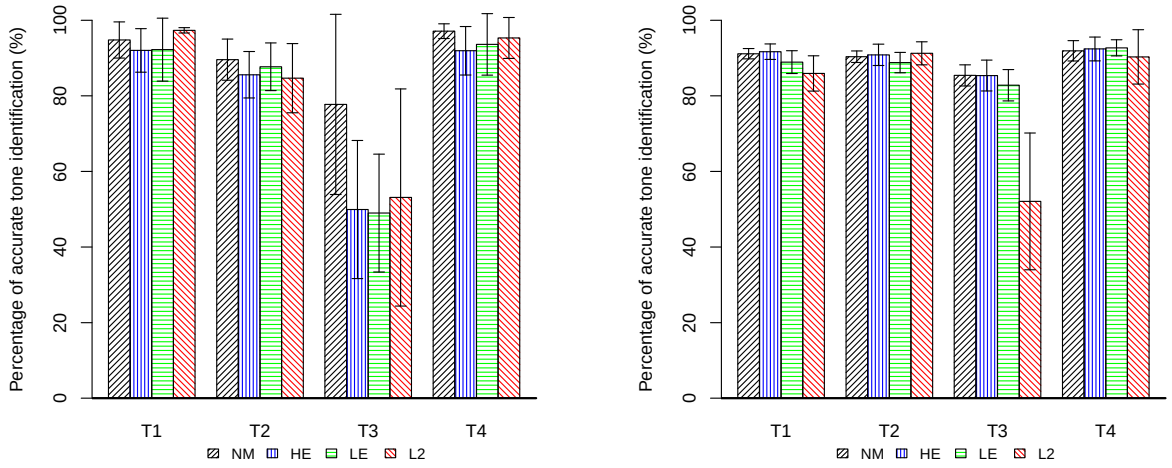
Note. Error bars indicate standard error of by-talker mean values.

The general pattern of durational shortening of T3 in multisyllabic items provided converging evidence of between-group differences in the production of T3. In particular, after controlling for speech rate (by the duration of T1 in multisyllabic items), the HL group showed shorter T3 durations than the L2 group in phrase-final position (β_{L2} = 0.37, t = 2.13, p_{MCMC} = .03; fixed effects: talker-specific contextual T1 mean duration, Sex, Group; random effects: Talker, Item); however, no significant L2-HL difference was found in non-phrase-final position, and no NM-HL difference was found in phrase-final or non-phrase-final positions (all $|t| < 1.4$, $p_{MCMC} > .5$). In short, with respect to both pitch contour and duration, HL speakers' reduction of T3 in non-phrase-final position more closely resembled that of NM speakers than did L2 speakers' production; however, HL speakers also showed a strong tendency to reduce T3 in phrase-final position, which was not found in NM or L2 speakers.

Perceptual Properties: Intelligibility, Goodness, and Sociolinguistic Classifiability

Figure 5 shows the mean likelihood of accurate tone identification (i.e., intelligibility score) for monosyllabic and multisyllabic items, based on tone and talker group. Generalized mixed-effects models of monosyllabic items' intelligibility scores showed that, overall, NM talkers' tones were more intelligible than HL talkers' (β_{NM} = 0.91, z = 3.17, p = .002), but HL talkers' tones were not generally more intelligible than L2 talkers' (p > .1). There was also a significant effect of Tone; in particular, T3 was the hardest to recognize across groups (β_{T3} = -1.87, z = -9.18, p < .001). However, T3 showed higher intelligibility in NM talkers' tokens than in HL talkers' ($\beta_{NM:T3}$ = 0.61, z = 5.26, p < .001). These results were consistent with the acoustic data discussed above, which showed both that HL and L2 talkers' tone production often differed from NM talkers' (especially in the production of T3), and that T3 was the most variable tone across groups. Further, the response times for HL speakers' tokens were longer than those for NM speakers' tokens (β_{NM} = -0.10, t = -2.80, p_{MCMC} = .01), and similar to the response times for L2 learners' tokens (β_{L2} = -0.002, t = -0.057, p_{MCMC} = .95). As expected, T3 tokens elicited the slowest response times (β_{T3} = 0.11, t = 4.88, p_{MCMC} < .001), although this effect was reduced for NM speakers' T3 tokens ($\beta_{NM:T3}$ = -0.062, t = -2.55, p_{MCMC} = .01).

Figure 5. Tonal intelligibility (i.e., tone identification accuracy) in (a) monosyllabic items and (b) multisyllabic items by tone and talker group, averaged over all talkers' mean values.

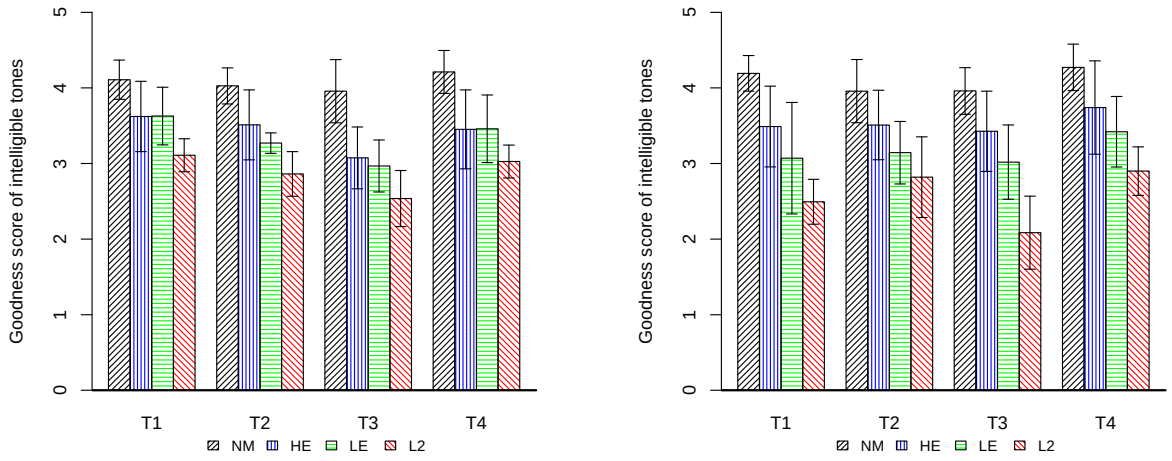


Note. Error bars indicate standard error of by-talker mean values.

As for the goodness of those isolated tones that were intelligible, as shown in Figure 6, HL speakers' tokens received goodness ratings that were lower than NM speakers' (β_{NM} = 0.71, t = 4.46, p_{MCMC} < .001), but higher than L2 learners' (β_{L2} = -0.47, t = -2.75, p_{MCMC} = .006). As with intelligibility, T3 received the lowest goodness ratings among the tones (β_{T3} = -0.38, t = -4.70, p_{MCMC} = .001), and this deficit was reduced in NM talkers' tokens ($\beta_{NM:T3}$ = 0.27, t = 6.03, p_{MCMC} < .001). In short, HL (both HE and LE) and L2 speakers' tones were more difficult to recognize

than those of NM speakers, but the perceived goodness of HL speakers' tones was in between that of the NM and L2 groups.

Figure 6. Goodness ratings for intelligible (i.e., correctly identified) tones in (a) monosyllabic and (b) multisyllabic items by tone and talker group, averaged over all talkers' mean values.



(a) (b)

Note. Error bars indicate standard error of by-talker mean values.

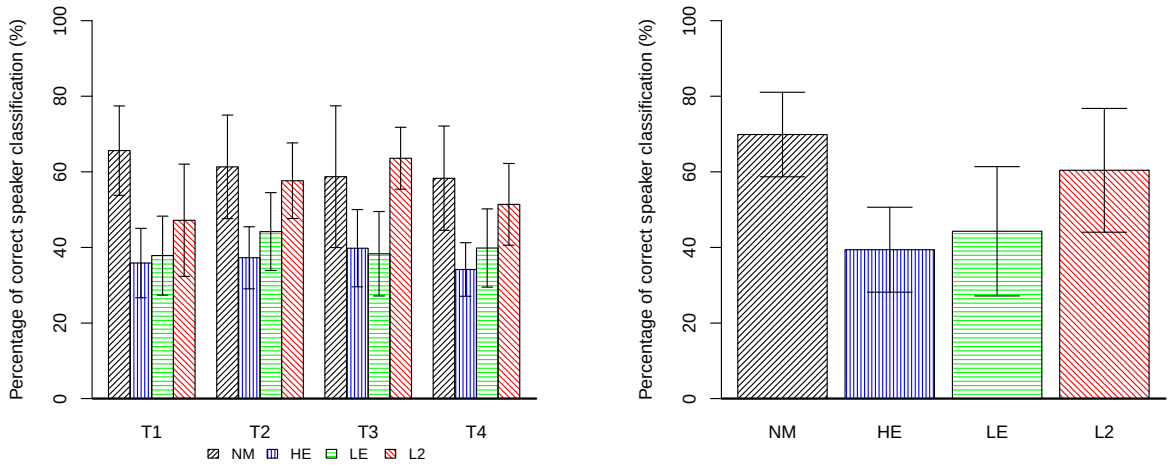
Tone identification results for multisyllabic items revealed some different patterns from the results for monosyllabic items. Compared to L2 speakers' tones, HL speakers' tones were easier to recognize (β_{L2} = -0.68, z = -4.14, p < .001), but not faster to evaluate. Compared to NM talkers', HL speakers' tones were slower to evaluate (β_{NM} = -0.064, t = -2.32, p_{MCMC} = .03), but there was no difference in intelligibility. As for goodness, HL speakers' tones again received intermediate ratings, which were lower than NM speakers' (β_{NM} = 0.70, t = 3.23, p_{MCMC} = .003) and higher than L2 speakers' (β_{L2} = -0.80, t = -3.43, p_{MCMC} < .001). The fact that HL speakers' tones were easier to identify than L2 speakers' in connected speech suggests that HL speakers had implemented connected-speech processes (e.g., tonal coarticulation, T3 reduction) in a more native-like manner, which is consistent with our acoustic results. In particular, Figure 3a showed how L2 speakers' failure to reduce T3 appropriately in non-phrase-final contexts resulted in T3 looking like T2 or T4. As in monosyllabic items, T3 showed lower intelligibility (β_{T3} = -0.43, z = -4.16, p < .001), lower goodness ratings (β_{T3} = -0.079, t = -2.30, p_{MCMC} = .02), and slower evaluation times (β_{T3} = 0.057, t = 3.29, p_{MCMC} = .01).

In regard to differences between the HE and LE groups, a comparison of listeners' judgments on HE and LE speakers' tokens (in terms of intelligibility, goodness ratings, and combined response times) yielded no significant difference for monosyllabic items, but two significant differences for multisyllabic items. HE speakers' tones in multisyllabic items were more intelligible overall, as shown by a Pearson's chi-squared test ($\chi^2(1, N = 5216) = 5.10$, p = .02), although the

difference in intelligibility rates was small (M_{HE} = 89.9%, M_{LE} = 87.9%); they also received higher goodness ratings ($t(4407.2) = 10.60$, $p < .001$; M_{HE} = 3.50, M_{LE} = 3.11).

Finally, with regard to sociolinguistic (demographic) classifiability, HL speakers emerged as the group that was the hardest to classify, as shown in Figure 7. On the basis of monosyllabic tokens, both NM and L2 speakers were more likely to be correctly classified than HL speakers (β_{NM} = 0.99, $z = 5.10$, $p < .001$; β_{L2} = 0.72, $z = 3.46$, $p < .001$). Listeners were also more confident about their classification of NM speakers than of HL speakers (β_{NM} = 0.25, $t = 3.88$, $p_{MCMC} = .001$); however, confidence ratings did not differ between the HL and L2 groups (β_{L2} = -0.0022, t = -0.003, $p_{MCMC} = .98$). Group had no effect on the combined response times for speaker classification and confidence rating with monosyllabic tokens (both $|t| < 1.9$, $p_{MCMC} > .06$).

Figure 7. Demographic classifiability (i.e., accuracy of identifying group affiliation) for (a) monosyllabic items by tone and group, and (b) multisyllabic items by group, averaged over all talkers' mean values.



(a) (b)

Note. Error bars indicate standard error of by-talker mean values.

In the case of multisyllabic tokens (which contained more acoustic and contextual information), HL speakers were still more difficult to classify than NM and L2 speakers (β_{NM} = 1.29, $z = 4.43$, $p < .001$; β_{L2} = 0.87, $z = 2.80$, $p = .005$). Moreover, listeners were the least confident in classifying HL speakers (β_{NM} = 0.20, $t = 2.23$, $p_{MCMC} = .04$; β_{L2} = 0.25, $t = 2.64$, $p_{MCMC} = .01$). The combined response times for classification and confidence ratings additionally showed that HL speakers took longer to classify than L2 learners, but not NM speakers (β_{NM} = -0.021, t = -0.32, $p_{MCMC} = .75$; β_{L2} = -0.18, t = -2.44, $p_{MCMC} = .01$). For both monosyllabic and multisyllabic items, no HE-LE difference was found in the confidence or (combined) reaction time of speaker classification; nevertheless, HE speakers were correctly classified at lower rates than LE speakers (monosyllabic items: M_{HE} = 36.5%, M_{LE} = 40.0%; multisyllabic items: M_{HE} = 39.4%, M_{LE} = 44.3%; Pearson's chi-squared tests yielded $\chi^2 > 6$, $p < .02$ for both contexts). In

other words, the difficulty of sociolinguistic classification was evident for both HE and LE speakers, but to a greater degree for HE speakers.

DISCUSSION AND CONCLUSIONS

In summary, the acoustic and perceptual data gathered in this study supported our four hypotheses regarding HL Mandarin speakers' tone production. First, the acoustic results showed a general pattern in which HL speakers did not uniformly resemble either NM or L2 speakers, but rather differed from one or the other group (or both) depending on the specific pattern or property. With respect to the durational shortening of T1 and T2 in multisyllabic contexts, HL speakers as a group patterned in between NM and L2 speakers. With respect to the turning point of T3, on the other hand, HL speakers resembled NM speakers. Along the same lines, with respect to T3 reduction in non-phrase-final contexts, HL speakers patterned distinctly from L2 speakers, and quite closely with NM speakers.

With respect to two other features, HL speakers again patterned distinctly from both NM and L2 speakers, but in this case they were located at the end, rather than in the middle, of the relevant continuum: of all groups, HL speakers produced the shortest T3 durations in phrase-final multisyllabic contexts, and showed the highest levels of pitch contour variability. As mentioned before, high tonal variability in the HL group, at least for the isolation forms of tones, would follow from the nature of HL speakers' experience with the target language, which does not consistently include the type of exposure to isolation forms received by native speakers (in the course of L1 education) and adult L2 learners (in the course of formal L2 instruction).[7] This educational disparity may also be related to HL speakers' relatively high rates of T3 reduction and concomitant durational shortening in phrase-final multisyllabic contexts, where they may not have not heard T3 pronounced in its "full" form as much as individuals exposed more to clear/emphatic speech in a regular (i.e., L1 or L2) classroom environment.

Our hypotheses about the perception of HL speakers' tones (and of their speech, more generally) were also supported overall. Taken together, the results for monosyllabic and multisyllabic items showed HL speakers' tones patterning not consistently like either NM or L2 speakers' tones in intelligibility: in monosyllabic items, they resembled L2 speakers', whereas in multisyllabic items they resembled NM speakers'. HL speakers' intelligible tones were also intermediate in native-likeness (i.e., between those of NM and L2 speakers). Furthermore, HL speakers were more difficult for native listeners to classify demographically than either NM or L2 speakers were. This last result suggests that, at least in the context of the demographic categories that listeners were given, HL speakers were the most ambiguous in terms of demographic background, which is consistent with their intermediate patterning in many of the acoustic properties discussed above.

Before we discuss the interpretations of these findings in further detail, it is important to acknowledge two limitations of this study. First, tone—like any phonological category—is multidimensional, and the manner in which we have parameterized it acoustically in this study represents only one of several ways in which tone could be analyzed. Although the numerous measures we have presented help to form a holistic picture of between-group differences in tone production, it remains possible that a different relative patterning of groups might be evident if

different metrics were considered. Second, although the perception study was focused on tone (and listeners were, therefore, instructed to rate the goodness of the tones, as opposed to the segments in each stimulus), we cannot guarantee that listeners' judgments were not influenced by the other characteristics of the speech they heard (e.g., voice onset time of stop consonants, vowel quality). Consequently, the perceived goodness data should be taken with the proverbial grain of salt, as they may not represent judgments of tone quality only.

Returning to the issue of high variability observed in the isolated tones of the HL group, we would like to point out that this variability may be related not only to the heterogeneity of HL speakers' experience with isolation forms, but also to the nature of their experience with regionally diverse varieties of Mandarin, which can show significant tonal differences from each other. For example, T3 in pre-pausal position is realized with a strictly falling contour (i.e., half T3) more often in Taiwan Mandarin and Singapore Mandarin than in Beijing Mandarin (Chua, 2003; Shih, 1988; Tai, 1978). Such tonal differences across varieties of Mandarin are relevant because of differences in the composition of the talker groups: whereas the majority of the NM group was from Mainland China, and most of the L2 group had experience primarily with a Mainland Mandarin variety, more than half of the HL group had at least some exposure to Taiwan Mandarin and/or Singapore Mandarin (see Chang, Yao, Haynes and Rhodes, 2011 for further details).

To explore the possibility that some of the increased variability in the HL group could be attributed to greater diversity in the HL group's Mandarin experience, we divided the NM, HE, and LE groups into subgroups according to whether their primary Mandarin experience was with a Mainland variety (ML) or with a southern variety—namely, Taiwan or Singapore Mandarin (TWSG)—and then conducted a post-hoc comparison of ML and TWSG talkers, taking as a test case rates of half T3 production in both phrase-final and non-phrase-final multisyllabic contexts.[8] This comparison revealed, in line with the literature on Chinese dialectology, a tendency for TWSG talkers to produce half T3 at higher rates than ML talkers; this was the case in every group and every context, with one exception: phrase-final position for the NM group, where TWSG talkers' rate of half T3 production was *lower* than that of ML talkers. Note that this reverse pattern observed for NM speakers in the U.S. makes it difficult to attribute the HL group's high rate of half T3 production phrase-finally (Figure 4) to greater TWSG exposure.

Moreover, while these results on T3 production lend credence to the idea that greater regional diversity in Mandarin experience may have increased the variability of T3 specifically, they do not account for the higher overall tonal variability observed in the HL group (which, as mentioned above, was not driven primarily by HL talkers' high variability on T3). Although we cannot say for sure that the HL group's high variability on the other tones does not also have a source in subtle dialectal variation, given that T3 production is described as one of the most salient loci of cross-dialect differences in tonal implementation, we consider it most likely for the HL group's high overall tonal variability to be due to the educational differences alluded to earlier: compared to educated NM speakers and instructed L2 learners, HL speakers without formal classroom experience in the HL tend to have little previous exposure to citation forms, so the tonal targets in a task eliciting citation forms are less well-defined for them. Thus, it bears repeating that the HL group's high tonal variability was specific to the isolation context (as

variability in multisyllabic contexts was not examined). Consequently, we are careful to point out that this variability does not necessarily reflect a production deficit; rather, it follows from the HL group's relative lack of experience with a particular register of the target language.

Whatever the cause of the HL group's tonal variability, this variability is likely to be a main contributor to the demographic ambiguity of the HL group observed in the sociolinguistic classification task. Although one could argue that HL speakers' demographic ambiguity is merely an artifact of the way the HL group was constituted (which resulted in the inclusion of a wider range of experience with the target language than in the NM and L2 groups), in many ways, this is exactly the point: the linguistic heterogeneity of HL Mandarin speakers makes it difficult to associate this population with a well-defined perceptual category. While some HL speakers may sound like native speakers, others sound more like L2 learners (and yet others, somewhere in between). It should, therefore, come as no surprise that native speakers are less adept at classifying HL speakers as *American-born Chinese* than they are at classifying native speakers and L2 learners as, respectively, *native Chinese* and *non-Chinese American*. Naturally, listeners could have varied with respect to their application of these demographic labels; it is possible, for example, that if a talker sounded like a 1.5-generation American (i.e., born in a Mandarin-speaking country, but raised in the U.S. from an early age), some listeners might have labeled this talker as *native Chinese* while others might have labeled the talker as *American-born Chinese*. Crucially, however, insofar as the labels used correspond to salient social categories for native Mandarin speakers, the lower degree of consistency in classifying HL talkers with one target label suggests that HL speakers, as a group, are perceived more variably than either native or L2 speakers.

In closing, our results point out several avenues of future research on HL phonetics and phonology. In the case of HL speakers of Mandarin, it remains unclear how HL speakers produce the neutral tone (which is known for being highly context-dependent in its phonetic realization), how their production of neutral tone compares to native and L2 speakers' production, and how the relative patterning of HL, native, and L2 groups on neutral tone compares to the relative patterning of these groups on the full tones. In addition, HL speakers' knowledge of other aspects of suprasegmental structure, such as intonation, requires systematic investigation. The contribution of the current results is in providing data from the suprasegmental domain that complement data from the segmental domain in showing that HL speakers' early experience with a target language can provide a measurable advantage over adult L2 learners in terms of approximating target norms, even if this advantage may not be clear in all aspects of speech production. These findings thus add to the growing body of evidence supporting the view that heritage speakers are language users distinct from both native and L2 speakers.

ACKNOWLEDGEMENTS
The authors are grateful to Sergio Infante, Chang Liu, and Jin Luo for research assistance and to Erin Haynes and Russell Rhomieux for their contributions to the larger research project associated with this study.

REFERENCES

Au, T. K., Knightly, L. M., Jun, S.-A., & Oh, J. S. (2002). Overhearing a language during childhood. *Psychological Science, 13*(3), 238–243.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Bates, D., Maechler, M., & Bolker, B. (2011). Lme4: Linear mixed-effects models using S4 classes [R package]. Version 0.999375-39. Available from http://cran.r-project.org/package=lme4

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer [Computer software]. Version 5.4.14. Available from http://www.praat.org

Chang, C. B. (2016). Bilingual perceptual benefits of experience with a heritage language. *Bilingualism: Language and Cognition, 19*(4), 791–809.

Chang, C. B., & Bowles, A. R. (2015). Context effects on second-language learning of tonal contrasts. *Journal of the Acoustical Society of America, 138*(6), 3703–3716.

Chang, C. [B.], & Yao, Y. (2007). Tone production in whispered Mandarin. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1085–1088). Dudweiler, Germany: Pirrot.

Chang, C. B., Yao, Y., Haynes, E. F., & Rhodes, R. (2011). Production of phonetic and phonological contrast by heritage speakers of Mandarin. *Journal of the Acoustical Society of America*, *129*(6), 3964–3980.

Chao, Y.-R. (1930). A system of tone-letters. *Le Maître Phonétique, 45*, 24–27.

Chao, Y.-R. (1933). Zhong guo zi diao gen yu diao [Tone and intonation in Mandarin Chinese]. *Guo Li Zhong Yang Yan Jiu Yuan Li Shi Yu Yan Yan Jiu Suo Ji Kan* [Journal of the Institute of History and Philosophy, Academia Sinica], *4*(2), 121–135.

Chua, C. L. (2003). *The emergence of Singapore Mandarin: A case study of language contact* (Unpublished doctoral dissertation). University of Wisconsin, Madison, Wisconsin.

Clopper, C. G. (2013). Modeling multi-level factors using linear mixed effects. *Proceedings of Meetings on Acoustics, 19*, 060028. Retrieved from http://dx.doi.org/10.1121/1.4799729

Dai, J.-h. E., & Zhang, L. (2008). What are the CHL learners inheriting? Habitus of the CHL learners. In A. W. He & Y. Xiao (Eds.), *Chinese as a heritage language: Fostering rooted world citizenry* (pp. 37–51). Honolulu, HI: National Foreign Language Resource Center, University of Hawaii.

Duanmu, S. (2007). *The phonology of standard Chinese*, 2nd ed. Oxford, England: Oxford University Press.

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics, 40*(2), 269–279.

Kiriloff, C. (1969). On the auditory perception of tones in Mandarin. *Phonetica, 20*(2–4), 63–67.

Knightly, L. M., Jun, S.-A., Oh, J. S., & Au, T. K. (2003). Production benefits of childhood overhearing. *Journal of the Acoustical Society of America, 114*(1), 465–474.

Kong, Y.-Y., & Zeng, F.-G. (2006). Temporal and spectral cues in Mandarin tone recognition. *Journal of the Acoustical Society of America, 120*(5), 2830–2840.

Kuang, J. (2013). *Phonation in tonal contrasts* (Unpublished doctoral dissertation). University of California, Los Angeles, California.

Lee-Ellis, S. (2012). *Looking into bilingualism through the heritage speaker's mind* (Unpublished doctoral dissertation). University of Maryland, College Park, Maryland.

Li, D., & Duff, P. A. (2008). Issues in Chinese heritage language education and research at the postsecondary level. In A. W. He & Y. Xiao (Eds.), *Chinese as a heritage language: Fostering rooted world citizenry* (pp. 13–32). Honolulu, HI: National Foreign Language Resource Center, University of Hawaii.

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech, 47*(2), 109–138.

Lukyanchenko, A., & Gor, K. (2011). Perceptual correlates of phonological representations in heritage speakers and L2 learners. In N. Danis, K. Mesh, & H. Sung (Eds.), *Proceedings of the 35th Annual Boston University Conference on Language Development*, vol. 2 (pp. 414–426). Somerville, MA: Cascadilla Press.

Montrul, S. (2002). Incomplete acquisition and attrition of Spanish tense/aspect distinctions in adult bilinguals. *Bilingualism: Language and Cognition, 5*(1), 39–68.

Montrul, S. (2004). Convergent outcomes in second language acquisition and first language loss. In B. Köpke, M. S. Schmid, M. Keijzer & L. Weilemar (Eds.), *First language attrition: Interdisciplinary perspectives on methodological issues* (pp. 259–280). Amsterdam, The Netherlands: John Benjamins.

Obrig, H., Rossi, S., Telkemeyer, S., & Wartenburger, I. (2010). From acoustic segmentation to language processing: Evidence from optical imaging. *Frontiers in Neuroenergetics, 2*(13), 1–12.

Oh, J. S., Au, T. K., & Jun, S.-A. (2010). Early childhood language memory in the speech perception of international adoptees. *Journal of Child Language, 37*(5), 1123–1132.

Oh, J. S., Jun, S.-A., Knightly, L. M., & Au, T. K. (2003). Holding on to childhood language memory. *Cognition, 86*(3), B53–B64.

Pallier, C., Dehaene, S., Poline, J.-B., LeBihan, D., Argenti, A.-M., Dupoux, E., & Mehler, J. (2003). Brain imaging of language plasticity in adopted adults: Can a second language replace the first? *Cerebral Cortex, 13*(2), 155–161.

Poeppel, D., & Hackl, M. (2008). The functional architecture of speech perception. In J. R. Pomerantz (Ed.), *Topics in integrative neuroscience: From cells to cognition* (pp. 154–180). Cambridge, England: Cambridge University Press.

Polinsky, M., & Kagan, O. (2007). Heritage languages: In the 'wild' and in the classroom. *Language and Linguistics Compass, 1*(5), 368–395.

R Development Core Team. (2015). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Available from http://www.r-project.org

Rao, R. (2015). Manifestations of /bdg/ in heritage speakers of Spanish. *Heritage Language Journal, 12*(1), 48–74.

Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech, 34*(2), 145–156.

Shen, X. S., Lin, M., & Yan, J. (1993). F0 turning point as an F0 cue to tonal contrast: A case study of Mandarin tones 2 and 3. *Journal of the Acoustical Society of America, 93*(4), 2241–2243.

Shi, F. (1986). Tianjin fangyan shuangzizu shengdiao fenxi [An analysis of the bisyllabic tones in Tianjin dialect]. *Yuyan Yanjiu, 1986.1,* 77–90.

Shi, F., & Wang, P. (2006). Beijinghua danziyin shengdiao de tongji fenxi [A statistical analysis of the tones in Beijing Mandarin]. *Zhongguo Yuwen, 2006.1,* 33–40.

Shih, C. (1988). Tone and intonation in Mandarin. *Working Papers of the Cornell Phonetics Laboratory, 3,* 83–109.

Tai, J. H.-Y. (1978). *Phonological changes in modern standard Chinese in the People's Republic of China since 1949*. Washington, DC: Office of Research, United States Information Agency.

Tsukada, K., Xu, H. L., & Xu Rattanasone, N. (2015). The perception of Mandarin lexical tones by listeners from different linguistic backgrounds. *Chinese as a Second Language Research, 4*(2), 141–161.

Ventureyra, V. A. G., Pallier, C., & Yoo, H.-Y. (2004). The loss of first language phonetic perception in adopted Koreans. *Journal of Neurolinguistics, 17*(1), 79–91.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*(1), 61–83.

Yang, B. (2015). *Perception and production of Mandarin tones by native speakers and L2 learners*. Berlin, Germany: Springer Verlag.

Zhu, X. (2004). Jipin guiyihua – ruhe chuli shengdiao de suiji chayi [F_0 normalization: How to deal with between-speaker tonal variations?]. *Yuyan Kexue* [Linguistic Science], *2004.3*(2), 3–19.

NOTES

1. See Polinsky and Kagan (2007) for arguments in favor of conceptualizing HL speakers in terms of the creole continuum of "basilectal", "mesolectal", and "acrolectal" varieties.

2. These facts are sometimes interpreted as supporting an alternative analysis of T3 in which the basic form is the "half" (as opposed to "full") allotone. Although in this study we assume that the "full" allotone is basic, note that this assumption is not crucial for our purposes. Our main goal is to describe the differences in tone production across groups and contexts, which can be accomplished just as easily under the alternative analysis of T3. Under the alternative analysis, for example, the pattern we describe below as T3 reduction in non-final contexts would simply be interpreted as T3 lengthening or enhancement in final contexts.

3. In fact, all L2 participants except one had at least two semesters of Mandarin exposure. The exception, whose duration of Mandarin exposure was 2.5 weeks, received this exposure in an immersion context, which probably amounts to more than the equivalent duration of regular college-level foreign language instruction. Since our results remained the same whether or not this participant was included in the analysis, we have reported findings on the full dataset.

4. Note that the label *American-born Chinese* for HL speakers is imperfect, because if taken at face value, it would not apply to all of the individuals in the HL group (some of whom were not actually born in the U.S.). However, this was the label chosen because it was likely to be familiar to the native Chinese judges. As pointed out by an anonymous reviewer, the nature of this label is social, so it is worth emphasizing that demographic judgments are indeed about talkers' socio-demographic background (as opposed to their language proficiency per se).

5. This is a little later than the turning point reported in Shi and Wang (2006) for Beijing Mandarin (around 25%). Note, however, that Shi and Wang selected sampling points on a different scale (9 time points from 0% to 100%, spaced at every 12.5%), so the difference between our results and theirs cannot be interpreted at face value. Nevertheless, it is fair to say that our measured turning point in T2 is later than that in Shi and Wang (2006).

6. An anonymous reviewer pointed out that the disparity in structural control of monosyllabic versus multisyllabic items could be problematic for our comparison of durations in these two contexts. Although this disparity is not ideal, overall, it probably strengthens the finding of durational shortening in multisyllabic contexts, because the way in which the multisyllabic items differ structurally from the monosyllabic items (in particular, the occurrence of voiced syllable onsets and codas only in the multisyllabic items) is likely to cause voiced interval durations in multisyllabic contexts to be relatively longer.

7. An anonymous reviewer pointed out that HL speakers often receive educational exposure to the HL in the context of weekend Chinese schools, and this was true of many (7 of 15) of our HL speakers. However, these speakers' descriptions, along with our own observations of Chinese Sunday schools in the Bay Area circa 2009, suggest that weekend Chinese classes are often taught by individuals without professional training in teaching Chinese, such that the mode of instruction may differ considerably from both typical L1 and typical L2 instruction in

a regular school setting. Furthermore, while some HL participants started these Chinese classes before the age of 6, others started much later, after receiving extensive HL exposure at home. Crucially, therefore, formal instruction on tones received by HL speakers (if any) is likely to differ both qualitatively and chronologically from that received by NM or L2 speakers.

8. Each subgroup contained at least two talkers. By group, the talker distribution across the ML and TWSG subgroups was, respectively, 4 versus 2 (NM), 2 versus 7 (HE), and 2 versus 3 (LE).