ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Speech Production of Mandarin Lexical Tones Among Canadian Elementary Students Enrolled in Mandarin–English Bilingual Schools

Youran Lin,[a] **Karen E. Pollock,**[a] **and Fangfang Li**[b]

[a] Department of Communication Sciences and Disorders, University of Alberta, Edmonton, Canada [b] Department of Psychology, University of Lethbridge, Alberta, Canada

ARTICLE INFO

Article History:
Received March 3, 2024
Revision received August 12, 2024
Accepted October 9, 2024

Editor-in-Chief: Cara E. Stepp
Editor: Leah Catherine Fabiano

https://doi.org/10.1044/2024_JSLHR-24-00150

ABSTRACT

Purpose: This study investigates how Mandarin–English bilingual students in Canada produce Mandarin tones and how this is influenced by factors such as tone complexity, cross-linguistic influences, and speech input.
Method: Participants were 82 students enrolled in a Chinese bilingual program in Western Canada. Students were recruited from Grades 1, 3, and 5 and divided into two groups based on their home language backgrounds: The heritage language group had early and strong input in Mandarin, and the second language (L2) group received mostly English input at home. Single-word tone productions were audio-recorded and transcribed by Mandarin-native listeners for match (accuracy) and pattern analyses. Acoustic measurements were extracted to provide phonetic details.
Results: First, Tone3 (dipping tone) was challenging across groups due to its complexity. Second, L2 students' productions were more influenced by English as a nontonal language and showed signs of categorical confusion. Third, increased tone match rates were related to both home input and school input, but bilingual students did not reach more than 90% of match rates in Grade 5. Instead, L2 students produced phonetic features less accurately in higher grades. This was attributed to reduced pronunciation instruction and limited home input.
Conclusions: Bilingual students' speech development in a minority language indicates unique influences of home and school input but also the universal influences of tone complexity. This study provides evidence for bilingual speech theories in the suprasegmental domain and has implications for the pedagogy of a minority language in the context of bilingual education.
Supplemental Material: https://doi.org/10.23641/asha.28098206

Bilingual speech development is often different from monolingual speech development due to (a) differences and interactions between two phonological systems and (b) varied quantity and quality of input in both languages (Baker & Trofimovich, 2005). Compared to the abundant research evidence on bilingual speech development in pairs of Indo-European languages (e.g., Goldstein et al., 2005; Menke, 2017; Nance, 2021; Sieg et al., 2023), less is known about how children acquire two phonological systems that are typologically different, for example, a tonal language and a nontonal language (Holm & Dodd, 2006; Kan & Schmid, 2019; Mok & Lee, 2018). Meanwhile, most studies on bilingual speech development have focused on an immigrant context where children of a language-minority group learn the speech system of the societal majority language (Munro & Derwing, 2020). Even less is known about how children with diverse language backgrounds acquire phonological skills in a minority language from school, where input at home and at school may play crucial roles in bilingual speech development.

The present study presents a unique population of school-aged children who are enrolled in a Mandarin–

Correspondence to Youran Lin: youran1@ualberta.ca. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

English bilingual program in Western Canada, including heritage speakers and second language (L2) learners of Mandarin. This study examines students' lexical tone productions in Mandarin, a phonological dimension that is nonexistent in English, and investigates the roles of home language backgrounds and schooling experiences. Results shed light on the learning of suprasegmental features in a societal minority language and provide evidence of relationships between home and school input. The following sections introduce the bilingual program as a testing field of bilingual speech development, review evidence on bilingual children's tone learning, summarize theories and factors of such learning, and pose specific research questions and hypotheses for the current study.

A Mandarin–English Two-Way Bilingual Program in Western Canada

In Canada, there are two major types of bilingual education programs: French immersion and two-way bilingual programs. French immersion programs usually involve a "one-way" immersion design, where students are L2 learners of French (nonfrancophone) and receive French input from their teachers, whereas francophone students may attend francophone schools under different programming (Canadian Charter, 1982; Dicks & Genesee, 2017). Evidence has shown that one-way immersion education was insufficient for L2 learners to acquire native-like pronunciation, despite an early age of onset, where students experienced "phonetic fossilization" (Netelenbos et al., 2016). This is attributed to the limited amount and reduced quality of L2 input in a foreign-language learning setting (Harada, 2007; Netelenbos et al., 2016).

In contrast, two-way bilingual programs provide instructions in both English and a minority language and enroll students from diverse backgrounds, including native and heritage speakers (Dicks & Genesee, 2017). The Mandarin–English two-way bilingual program in the present study enrolls students who recently immigrated to Canada, who were born in Canada and speak Mandarin or other Chinese languages at home, and who learn Mandarin as an L2 (Lin et al., 2024). At the elementary level (Grades 1–6), half of the class content is delivered in Mandarin Chinese, and the other half in English. Students are expected to develop or maintain functional proficiency in Mandarin at no cost to their English skills (Alberta Education, 2006).

The two-way design is considered to be of merit as it provides authentic input from native-speaking teachers and peers of both languages (Cummins, 1979). For example, Menke (2015, 2017) examined the vowel productions of Spanish-L2 students in a one-way Spanish immersion school and a two-way Spanish–English school in the United States, respectively, in Grades 1, 3, 5, and 7. One-way immersion students' vowel formants were different from those of Spanish-native peers, and such differences were larger in higher grade levels (Menke, 2015). As for two-way immersion students, little difference was detected from Spanish-native peers as early as in Grade 1 (Menke, 2017). The more native-like learning outcomes among the latter Spanish learners were associated with their regular contact with Spanish-native peers through the two-way bilingual program. Different from the Spanish students who often still had familial connections or ambient exposure to Spanish outside of school, Nance (2021) compared pre-adolescent (ages 7–11 years) students who had little to no exposure to Gaelic versus students who had home input in Gaelic, both enrolled in a Gaelic–English two-way bilingual program in Scotland. No phonological or acoustic difference was detected in their consonant productions. The author concluded that any initial differences in home language background had been leveled out by age 7 years, as the students had received enough input from the two-way bilingual education through peer interaction.

With such evidence, one might expect the speech production of Mandarin-L2 learners to converge with, or "catch up" (Nance, 2021, p. 370) to, their Mandarin-native or Mandarin-heritage peers through two-way bilingual education, despite the initial differences related to home language backgrounds. However, teachers of the Mandarin program reported that students had difficulties learning Mandarin tones, which persisted in higher grades (Lin et al., 2024). That is to say, gaps exist between learning outcomes as perceived by teachers and the belief in the effectiveness of two-way bilingual education. Such gaps might be related to the uniqueness of Mandarin tones and limited input in Mandarin as a minority language. Therefore, the Mandarin bilingual program provides a testing field to understand bilingual speech development and its relationships with bilingual phonological systems and bilingual speech input.

Evidence on Bilingual Children's Tone Development

Research is scarce on tone learning in bilingual education contexts (except for Meckelborg et al., 2024, which focuses on the effect of tonal language background on learning another tonal language for new learners), but evidence on children who maintain a tonal language as their heritage language (HL) can inform the current study. Studies on bilingual children's Mandarin tone learning mostly focused on perception, and very few touched on production. School-aged Mandarin–English bilingual children exhibited categorical perceptions of tones similar to monolingual peers and as opposed to the continuous

perception of English monolingual children (J. Yang & Liu, 2012) and showed similar skills of tone discrimination to monolingual peers (Marinova-Todd et al., 2010). Nonetheless, for Cantonese, Kan and Schmid (2019) found that school-aged (ages 5–11 years) Cantonese–English bilingual children in the United States scored lower than monolingual peers in tone discrimination. However, accurate categorization does not necessarily imply accurate production of phonetic characteristics. It remains unclear how school-aged children develop their tone productions in Mandarin.

Relevant research evidence on bilingual children's tone productions is available in Cantonese and shows mixed results. On one hand, research documented similar development between bilingual and monolingual children. Holm and Dodd (2006) found young Cantonese–first language (L1) children in Australia (ages 2;0–5;7 [years; months]) who sequentially learned English had similar tone match (accuracy) rates compared with monolingual peers in Hong Kong. Mok and Lee (2018) found simultaneous Cantonese–English bilingual children in Hong Kong (ages 2;0–2;6) had comparable tone match rates with monolingual peers, and the two rising tones with similar fundamental frequency (*F*0) contours were challenging despite language backgrounds. On the other hand, different developmental patterns between bilingual children and monolingual peers were noted in some studies. In Mok and Lee, bilingual children showed a high-low–pitch template in disyllabic tone productions, indicating the influences of English trochaic stress patterns. Yao et al. (2020) found that Urdu–Cantonese bilingual children in Hong Kong (ages 4;5–6;6) were more prone to tone mismatches than monolingual peers. Mok and Lee advocated for more studies involving a variety of tonal languages, including Mandarin, to better understand bilingual speech development in the suprasegmental domain.

The varied results in bilingual children's Cantonese tone production may be due to the variety of factors of bilingual speech learning, such as specific tone targets (e.g., phonetically similar rising tones vs. phonetically distinguishable tones), L1 transfer (e.g., Urdu-L1 vs. English-L1), and language environments (e.g., Australia vs. Hong Kong). These factors are discussed in bilingual speech development theories and are reviewed in the next sections.

Bilingual Speech Development Theories That Are Applicable to Bilingual Tone Production

There are several theoretical frameworks to account for bilingual speech development, for example, the Perceptual Assimilation Model (PAM; Best & Tyler, 2007) and the Speech Learning Model–Revised (SLM-r; Flege & Bohn, 2021). Both theories discuss interactions between learners' two phonological systems. Specifically, they address how learners' perception is attuned by their L1, which influences the perception and production of L2 phonological contrasts (Best & Tyler, 2007) and phonetic categories (Flege & Bohn, 2021). However, these theories primarily focus on speech sounds and provide limited accounts for suprasegmental categories (see, however, So & Best, 2010). In particular, there has been limited discussion on how children in a nontonal language environment learn tones in L2. Lexical tones are critical phonological categories in tonal languages, where pitch and other suprasegmental features are used to contrast meanings. Despite their early acquisition in native speakers (Holm & Dodd, 2006; Zhu, 2002), tones are challenging for L2 learners with a nontonal background. In a nontonal language, infants learn to ignore suprasegmental information in word recognition for efficient lexical processing (Singh et al., 2008). Therefore, nontonal adult L2 learners' tone productions are prone to errors, and their tonal speech can be harder to understand (Hao, 2012; C. Yang, 2016). Research on English-speaking children's production of Mandarin tones is needed to understand the ongoing process of speech development among younger learners in tonal–nontonal L1–L2 pairs and provide evidence for L2 speech theories in the suprasegmental domain.

In addition to L1–L2 interactions, L2 speech learning theories stress the roles of L2 input. Such input has quantitative and qualitative dimensions (Flege & Bohn, 2021) and may differ across various social contexts (Bedore et al., 2016). There is reason to believe that early input, usually in the home environment, shapes learners' perception (Best & Tyler, 2007). However, although research indicates that adult heritage speakers have advantages in speech production compared to L2 learners (Chang et al., 2011), for young immigrant children, their language dominance shifts to the societal majority language quickly, and their speech production may experience attrition and become similar to L2 learners with limited HL exposure (J. Yang & Fox, 2017). On the other hand, it is argued that L2 learning is a lifelong process influenced by recent input (Flege & Bohn, 2021), for example, the immersive input from a bilingual school. Evidence indicates school-aged L2 learners can catch up with their HL peers if they currently receive intensive, high-quality L2 input from school (Menke, 2017; Nance, 2021). Nonetheless, the evidence was from programs with Indo-European language pairs (Spanish–English and Gaelic–English programs, respectively), while no exploration has been made when the two languages of instruction are typologically distant and the L1–L2 interactions may be less facilitative. Therefore, the present study looked into a Mandarin–English bilingual program to understand the unique effects of L1–L2 interaction and home versus school input.

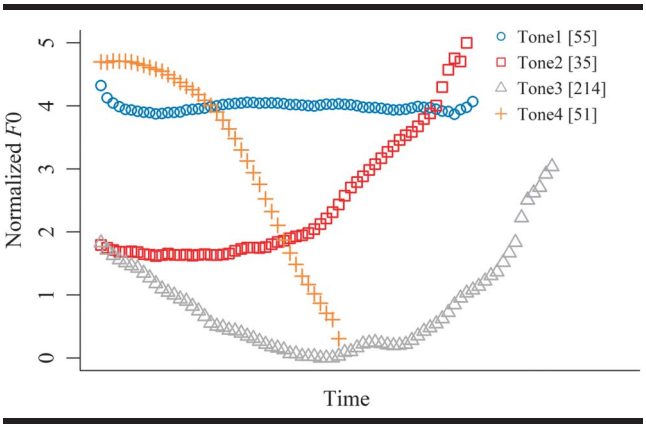
Factors Affecting Bilingual Children's Learning of Mandarin Tones

Synthesizing the aforementioned evidence and theories, bilingual children's tone learning may be impacted by three levels of factors: (a) intralanguage factors, that is, phonemic and phonetic complexity of the targets; (b) interlanguage factors, that is, transfer effects between L1 and L2; and (c) extralanguage factors, which may include child internal factors such as cognitive abilities and socioemotional development and child external factors such as the quantity and quality of bilingual input in the child's language experiences across different periods of time and social contexts (Bedore et al., 2016; Paradis, 2023). For extralanguage factors, the present study is particularly interested in Canadian Mandarin learners' speech input at home and at school.

Intralanguage Factors: Phonemic and Phonetic Complexities of Tone Targets

Evidence in monolingual tone acquisition provides insights into the complexity of tone targets. Mandarin has four citation tones in monosyllables: a high-level tone (Tone1), a mid-rising tone (Tone2), a low-dipping or falling–rising tone (Tone3), and a high-falling tone (Tone4). These can be transcribed in the five-scale convention (Chao, 1930): Tone1 [55], Tone2 [35], Tone3 [214], and Tone4 [51], where larger numbers indicate higher pitch in the speaker's tonal pitch range (see Figure 1). In addition, Mandarin tones are produced with different durations: Tone3 is usually produced with the longest duration, and Tone4 is the shortest (C. Yang, 2016). Furthermore, Tone3 [214] often co-occurs with a creaky voice (Kuang, 2017). All these acoustic characteristics were shown to contribute to listeners' tone perception (Blicher et al., 1990; Rhee et al., 2020). Due to space limits, this study focuses on pitch and duration in tone productions.

Figure 1. Mandarin citation tones produced by a female native speaker. Fundamental frequency (*F*0) was normalized using Equation 1.



In terms of the age of establishment, Mandarin citation tones as a whole category are established at 1.5 years of age in monolingual children (Zhu, 2002). Zhu (2002) believed that tone contrasts were mastered early because they were phonologically and perceptually salient. However, when examined through acoustic analyses, monolingual children's tone productions were not yet adultlike at the age of 5 years. Such acoustic differences existed despite the use of normalization methods to account for age differences (Xu et al., 2018) and were related to lower match rates (Wong, 2012). Some researchers believed that the protracted phonetic refinement was related to the development of muscle physiology and motor skills in young children (Wong, 2012), and others believed that it was a process of integrating phonetic cues (Rhee et al., 2020). In either case, the protracted refinement of tone productions indicated the complexity of Mandarin tones as phonetic targets, despite the early development of the phonemic categories. Therefore, the present study integrated both phonological and acoustic measures.

In terms of developmental order, Tone1 and Tone4 are acquired early by monolingual children, and Tone2 and Tone3 are acquired relatively later (Wong, 2012; Zhu, 2002). Evidence in adult L2 learners is comparable, suggesting that Tone3 productions had the lowest match rate and were frequently perceived as Tone2 (Wang et al., 2003; C. Yang, 2016). The challenge of producing Tone3 can be attributed to phonetic and phonological reasons. First, Tone3 is difficult to articulate with its compound *F*0 contour and extremely low pitch (Wong, 2012). Second, Tone3 is confusable with Tone2 because both involve a rising section (Blicher et al., 1990). Third, Tone3's allophonic realization is complex. It experiences a full sandhi when followed by another Tone3, realizing as a rising tone that is identical to Tone2 [35], and it experiences semi-sandhi when followed by a non-Tone3 syllable and becomes a short, low-falling tone [21] (Xu et al., 2018; C. Yang, 2016). In terms of the other later-developing tone, Tone2, Wang et al. (2003) showed that Tone2 was responsive to perceptual training among adult L2 learners. Therefore, the current study hypothesizes that in bilingual students' tone productions, Tone3 has the lowest match rate, and Tone1, Tone2, and Tone4 have similarly moderate to high match rates.

Interlanguage Factors: Cross-Linguistic Influences

For bilingual learners, another level of complexity of tone learning is cross-linguistic influences. Among L2 speech theories, PAM made the most explicit effort to specify L1 influences in the suprasegmental domain (PAM-S; So & Best, 2010). PAM-S proposed two

hypotheses of English-L1 listeners' perception of Mandarin tones. On one hand, tones may be perceived as prosodic categories in English, such as intonations or stress patterns. For example, Tone2 [35] may be assimilated to question intonation or iambic stress patterns, and Tone4 [51] may be assimilated to statement intonation or trochaic stress patterns (Hallé et al., 2004; So & Best, 2010). On the other hand, tones may be perceived as nonlinguistic melodies (Hallé et al., 2004), and the learning depends on their phonetic properties. Consequently, tones that share phonetic similarities are more confusable, for example, Tone1–Tone4 ([55]–[51]), Tone1–Tone2 ([55]–[35]), and Tone2–Tone3 ([35]–[214]; So & Best, 2010). Production studies confirmed that Tone3 was often produced as Tone2 by L2 learners (Wang et al., 2003). Based on such evidence, it is hypothesized that students with an English home background do not assign as much phonemic significance to tones due to English influences, so their productions have lower match rates. It is also hypothesized that tone pairs that can be assimilated into different prosodic categories in English are produced with less confusion (e.g., Tone2–Tone4), whereas pairs that are phonetically similar are subjected to substitution (e.g., Tone2–Tone3).

Extralanguage Factors: Effects of Speech Input at Home and at School

Different from PAM, which focuses on L1 assimilation, SLM-r (Flege & Bohn, 2021) highlights the role of L2 input. SLM-r views speech learning as a dynamic process that is a function of the bilingual speech input. Although SLM-r did not provide an operational definition of speech input, previous studies on language development suggested examining input across different periods of time and social contexts (Bedore et al., 2016; Paradis, 2023). Specific to the current study, the language of interest, Mandarin, is a minority language in Canada, which means the acquisition of phonological skills in Mandarin relies on the input at home and at school. Therefore, in this study, input was operationalized as the students' home language backgrounds (early and ongoing home input) and grade levels (school input). In the Mandarin program, Mandarin-L2 students' tone accuracy was lower than that of heritage peers in Grade 1, when tones were first introduced to them, indicating the effect of home input (Meckelborg et al., 2024). It is particularly of interest whether these Mandarin-L2 learners are able to acquire phonological skills in a minority language through schooling, as the goal of bilingual education is for students with diverse language backgrounds to develop functional bilingualism (Alberta Education, 2006). Evidence from two-way bilingual programs of Indo-European languages suggested that bilingual education provided high-quality input in the minority languages from teachers and native-speaking peers, and therefore, heritage speakers and L2 learners exhibited no articulation differences after a short period of immersion (Menke, 2017; Nance, 2021). Based on the existing research evidence, the current study hypothesizes an interaction effect between home language and grade level. That is, in lower grade levels, students with more home input (i.e., heritage learners) will outperform their L2-learning peers, but as school input accumulates (i.e., in higher grade levels), Mandarin-L2 learners will catch up to their heritage peers. Nonetheless, as a reminder, this does not align with teachers' observation (Lin et al., 2024).

The Current Study

This study investigates the development of Mandarin lexical tones in students who are enrolled in a Mandarin–English two-way bilingual program in Western Canada. Both transcription-based analyses (match rates and mismatch patterns) and acoustic analyses (pitch contours and duration) are used to examine tone productions. Research questions are guided by the multifaceted factors reviewed above: (a) In terms of intralanguage factors, which specific tone target(s) is the most challenging for bilingual students to produce? (b) In terms of interlanguage factors, do bilingual students' tone productions indicate cross-linguistic influences from English? (c) In terms of extralanguage factors, what are the impacts of home language backgrounds and schooling experiences? As a reminder, the hypotheses are as follows: (a) Tone3 is the most challenging tone target. (b) L2 learners of Mandarin (with English home language background) produce tones with lower match rates and more confusion between tone pairs that are phonetically similar (e.g., Tone2–Tone3), while tone pairs that can be assimilated into different prosodic categories in English are produced with less confusion (e.g., Tone2–Tone4). (c) More home and school input in Mandarin is associated with better performance in tone productions, and the home language differences can be leveled out through schooling.

To our knowledge, little research has examined Mandarin tone productions in school-aged children in a context where Mandarin is a minority language (Meckelborg et al., 2024). Thus, the current study will not only provide evidence for L2 speech learning theories in the suprasegmental domain but also have evidence-based implications for pronunciation teaching and learning in bilingual education.

Method

Participants

The study received ethics approval from the University of Alberta Research Ethics Board (No. Pro00075638),

and participants provided informed consent before taking part. Students in Grades 1, 3, and 5 at Mandarin bilingual schools were voluntarily registered by their parents. In total, 165 students were recruited for a larger project. In the present study, 82 students were selected based on their parent-reported language exposure and use at home to represent two distinct background groups: 38 were HL speakers of Mandarin (group name "HL") with early and strong Mandarin home input, and 44 were L2 learners (group name "L2") with mostly English home input and late onset of Mandarin. Parents reported either no diagnoses of hearing, speech, language, or learning problems or previous diagnoses of mild disorders prekindergarten (N = 5). All but two participants passed a hearing screening. The two participants who did not pass the screening failed at one or two frequencies in one ear and were referred to their pediatricians for follow-up. Since no concerns about hearing were reported, these children's speech samples were included in the study.

The two groups' profiles are depicted in Table 1. Participant numbers were balanced across grade levels. Even within the L2 group, there were often extended family members who spoke Mandarin or another Chinese language. Within the HL group, students had various onset of English exposure (M = 28 months, SD = 26), and most students spoke English as their current dominant language. A few students were from mixed families where they had a small amount of exposure to non–Mandarin Chinese varieties such as Cantonese, Teochew, Hakka, and Toishanese, or a non-Chinese language such as Tagalog, from one of their parents or grandparents. However, it is clear that the two groups had different home language environments: HL students had early onset of exposure to Mandarin (0–1 month of age), whereas L2 students had late onset of Mandarin (older than 3 years of age) and early onset of English (0–1 month of age); HL students had at least one parent who spoke Mandarin ≥ 50% of the time, whereas L2 students had both parents who spoke English ≥ 80% of the time; and HL students had at least one parent who self-reported high proficiency in Mandarin (4 or 5 out of the 0–5 scale), whereas L2 students had both parents with high proficiency in English and low proficiency in Mandarin. As a result, HL students had higher proficiency in Mandarin than L2 students, indicated by raw scores of the Chinese Peabody Picture Vocabulary Test (Lu & Liu, 1998). Note that the exposure to Mandarin was based on parent report, which could include varieties spoken in areas such as Mainland China, Taiwan, and Malaysia as suggested by parents' birthplaces.

In addition, 12 Chinese teachers in the same bilingual program provided speech samples. Among them, seven were L1 speakers of Mandarin, five were L1 speakers of non-Mandarin varieties of Chinese and started

Table 1. Demographic information and home language environment of heritage language (HL) and second language (L2) students (M (SD) [range] for numeric measurements and median [range] for ordinal measurements).

Information	HL	L2
Participant numbers	G1 = 15	G1 = 16
	G3 = 11	G3 = 14
	G5 = 12	G5 = 14
Chronological age in months	G1: 77 (3) [72, 83]	G1: 78 (4) [71, 86]
	G3: 104 (4) [97, 110]	G3: 103 (4) [97, 111]
	G5: 124 (3) [120, 129]	G5: 126 (3) [122, 130]
Percentage of students whose current dominant language is English	G1: 53; G3: 82; G5: 83	100
Onset age of regular exposure to Mandarin (months)[a]	0 (0) [0, 1]	63 (11) [37, 96]
Onset age of regular exposure to English (months)[a]	28 (26) [0, 93]	0 (0) [0, 1]
Parent percentage of time speaking Mandarin to the child[b]	72 (33) [0, 100]	0 (0) [0, 1]
	89 (13) [50, 100]	2 (4) [0, 15]
Parent percentage time speaking English to the child[b]	10 (14) [0, 50]	96 (7) [80, 100]
	22 (28) [0, 100]	99 (2) [90, 100]
Parent self-reported Mandarin proficiency (scale 0–5)[b]	5 [0, 5]	0 [0, 1]
	5 [4, 5]	1 [0, 3]
Parent self-reported English proficiency (scale 0–5)[b]	3 [0, 4]	5 [4, 5]
	3 [2, 5]	5 [5, 5]
CPPVT scores (full mark = 99)	55 (21) [18, 83]	11 (6) [0, 25]

Note. G = Grade; CPPVT = Chinese Peabody Picture Vocabulary Test.

[a]When parents did not report an onset for the exposure to a language at home, it was assumed that English exposure started at 60 months (kindergarten) and Mandarin started at 72 months (Grade 1). [b]The data of two parents (when applicable) are presented in the order of the smaller number and the larger number between the two parents.

learning Mandarin in elementary school, and two were born in Canada and graduated from the bilingual program. These teachers formed a representative sample to understand the Mandarin input students receive at school.

Procedure

Questionnaire. Parents filled out a questionnaire to quantify their children's language environment and experiences. It was adapted from the Language Experience and Proficiency Questionnaire (Marian et al., 2007) and the Alberta Language Environment Questionnaire (Paradis, 2011) and made available in both English and Chinese.

Speech samples. A picture-based single-word elicitation test was adapted from Zhao and Bernhardt (2012) and Zhu (2002). The test included 72 target words. Among them, 32 were monosyllabic targets, which were the focus of the current study (see the Appendix for the list). Examples of eliciting questions included, "What is this?" and "What is this person doing?" No written prompt was provided. Three examiners, who were Mandarin-L1 speakers and proficient in English, elicited productions that were as spontaneous as possible, but imitative models were provided as needed. Speech samples were audio-recorded using a Zoom H1n digital recorder with a Pro Lavalier JK MIC-J 055 unidirectional cardioid condenser microphone positioned in front of the child's chest. The recordings were mono audios at a 48-kHz sampling rate and 24-bit resolution.

Phonetic transcription. Speech samples were transcribed by four Mandarin-L1 researchers. Transcribers coded whether each production was spontaneous or imitative. Subsequently, each tone production was transcribed as one of the four citation tones; the semi-sandhi of Tone3 [21], which is inappropriate in the monosyllabic context (Xu et al., 2018); or an uncategorizable production. In the whole word list, 23% of the samples were transcribed by a second transcriber and reached 90% intertranscriber reliability. This was interpreted as acceptable since allophonic variations and uncategorizable productions were considered (Shriberg et al., 1997). The first transcribers' transcriptions were adopted for analysis.

Spontaneity. Both groups of students made a considerable number of imitative productions. Imitative models can significantly increase tone match in bilingual students (L. Yang et al., 2021). Moreover, a Mann–Whitney U test showed a significant group effect on the number of spontaneous productions ($U = 105.5$, $p < .001$): The L2 group ($M = 18.386$, $SD = 6.721$) tended to produce fewer spontaneous tones than HL ($M = 29.763$, $SD = 4.863$), which will compound the results. Therefore, only spontaneous productions were analyzed. This left 1,131 productions by HL students, 809 by L2 students, and 387 by teachers (all spontaneous).

Tone match analysis. In this study, tone analysis was fulfilled in Phon (Hedlund & Rose, 2020). A new functionality was developed to compare the transcribed tone against the target on each syllable. A full list of individual word productions can be generated with detailed information about the speakers, spontaneity, target words, target tones, and transcriber perceived tones to support token-level match analyses and mismatch pattern analyses. Tone substitution, semi-Tone3, and the production of uncategorizable tones were all coded as mismatch.

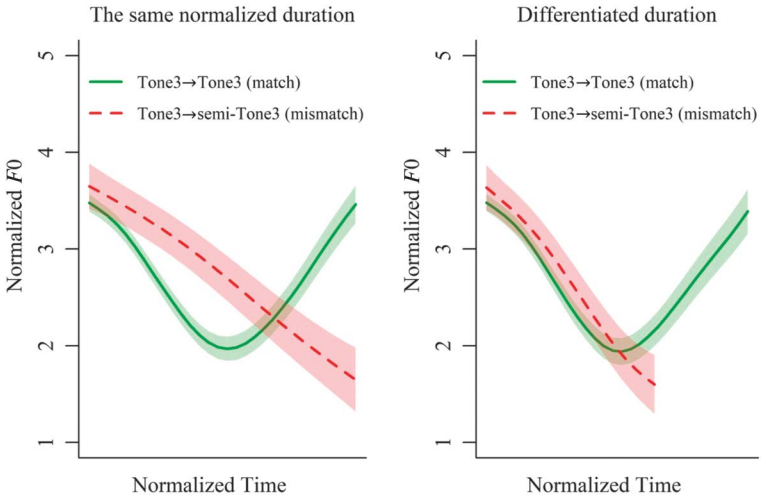
Acoustic measurement extraction. Tones were labeled over the nucleus vowel and any voiced segment after it (Wong, 2012; Xu, 1997) in Praat (Boersma & Weenink, 2023). The onset of a vowel after a voiceless consonant was identified through cues such as regular voice pulses, a shift in formants, a drop in intensity, and increased regularity in the waveform. The end of a tone was marked at the last zero-crossing point with a clear $F0$ contour and formant structure. Four cycles at the beginning and at the end were excluded to eliminate irregular pitch patterns (Wong, 2012). ProsodyPro was used to extract duration (milliseconds) and $F0$ (Hertz), where automatically recognized voicing pulses were examined and manually adjusted (Xu, 2013). Ten $F0$ values with equal time intervals were extracted for each syllable.

$F0$ normalization. The $F0$ values were normalized into T values within each speaker according to Equation 1 (Shi & Wang, 2006). This equation compresses $F0$ differences in the higher pitch range through a logarithmic transformation, scales a speaker's $F0$ by their own range, and converts the logarithmic values into a 0–5 scale to be comparable with Chao's (1930) 5-scale tone letters. This method is able to minimize anatomical variation and sociolinguistic variation across speakers and preserve the phonemic distinctions between tone categories (Zhang, 2018).

$$T = 5 \times \frac{\log_{10}(F0) - \log_{10}(F0_{\min})}{\log_{10}(F0_{\max}) - \log_{10}(F0_{\min})} \quad (1)$$

Differentiated durations. When $F0$ contours are plotted with the same normalized time, the contours can be misleading. For example, in Figure 2, some of the Tone3 are transcribed as a match (i.e., perceived as [214]) and the others as a semi-Tone3 [21]. When the two types of productions are plotted with equal durations, they barely overlap. However, when their average durations are considered, semi-Tone3 overlaps with the falling section of Tone3, which is a better reflection of the phonetic reality. Therefore, the duration of $F0$ contours was differentiated

Figure 2. Plots of Tone3→Tone3 productions and Tone3→semi-Tone3 productions with equal durations and differentiated durations. *F*0 = fundamental frequency.



by the average of subgroups, defined by target tone, transcribed tone, grade, and background (see Supplemental Material S1).

Statistical analysis. Statistical analyses were conducted in R Version 4.2.3 (R Core Team, 2023). Logistic mixed models of tone match were conducted using the glmer() function in the lme4 package (Bates et al., 2015). Generalized additive mixed models (GAMMs) of *F*0 contours were conducted using mgcv and itsadug in R (van Rij et al., 2022; Wood, 2017). Mixed linear models of duration were conducted using the lmer() function in the lme4 package.

Results

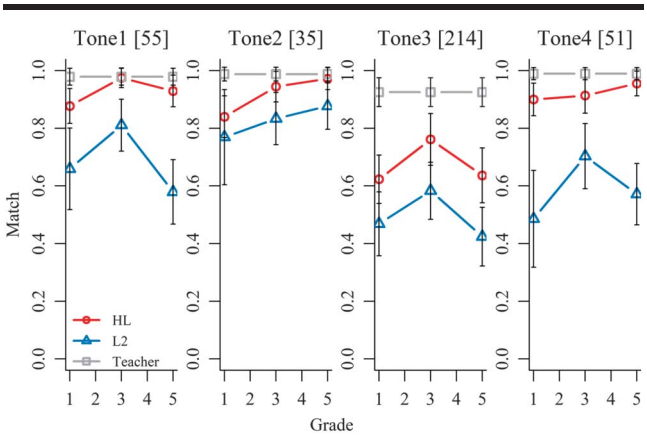
Since the study involves multiple analyses (transcription-based analyses such as match rates and mismatch patterns and acoustic analyses of *F*0 contours and duration) and each research question may be addressed in multiple analyses, brief interpretations and summaries of results are provided when appropriate. Subsequently, the Discussion section synthesizes the results and addresses the research questions that focus on the multifaceted factors of bilingual students' tone productions.

Tone Match (Accuracy)

Match analyses can help understand bilingual students' tone learning outcomes and the factors that might have impacted such learning. Students' and teachers' match rates of spontaneously produced citation tones were plotted by tone target, group, and grade (see Figure 3). These plots of raw data suggest that Tone3 was the most challenging for all students and even for teachers. HL students' tone match rates were similar to teachers', except for Tone3, which had lower match rates. On the other hand, L2 students' match rates were generally lower than those of teachers and HL students, except for Tone2, which was comparable with HL students. Both HL and L2 students achieved high match rates in Grade 3, but there seemed to be a trend for L2 students' match to be lower in Grade 5, whereas HL students seemed resistant to this trend.

A logistic mixed model was built to confirm the observed patterns. The target was a binomial variable of the match of each production (match or mismatch). Fixed effects included group (HL and L2), grade level (Grades 1, 3, and 5), and tone target (Tones 1, 2, 3, and 4), and

Figure 3. Raw data of average match rates of different tones by group and grade with teachers' data as a reference across grade levels. Error bars mark 95% confidence intervals. HL = heritage language; L2 = second language.

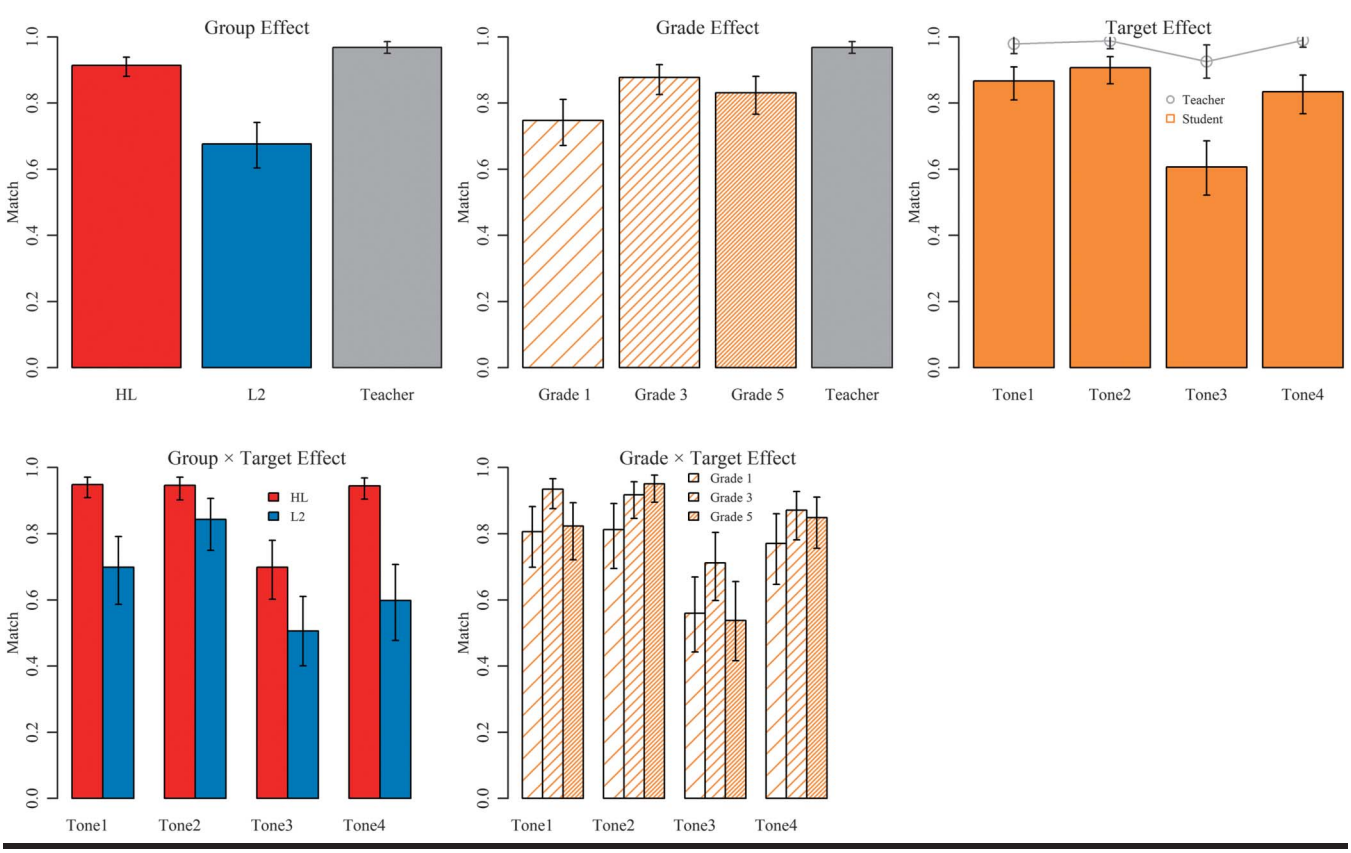


simple contrast was used. Random effects included target word and speaker. Teacher productions were not included in the model because they did not have the dimensions of grade or group. Fixed effects and interactions were added to the model in a stepwise manner, and Akaike information criterion (AIC; criterion = 2) was used to compare models. The selected optimal model included fixed effects of group, grade, target, and interactions of Target × Group and Target × Grade (see Supplemental Material S2). The fixed effects aforementioned were presented in Figure 4, with raw data of teachers' match rates plotted as a reference when appropriate. The *p* values were Bonferroni adjusted for contrasts of interest. A threshold value of probability $\alpha = .05$ was used to evaluate the significant difference, $\alpha = .1$ was used to evaluate the marginal difference, and $p > .1$ indicated no significant difference. Between-groups comparisons were achieved using emmeans, and the mean differences (*MD*s) reported below represent the differences of estimated probabilities of match (being accurate), converted using the plogis() function.

The model indicated the effect of language background. The HL group had a significantly higher probability of tone match than L2 ($MD = 0.238$, $z = 7.205$, $p < .001$). Furthermore, the model indicated the effect of grade level. The difference was significant between Grade 1 and Grade 3 students, with a higher probability of match in Grade 3 ($MD = 0.130$, $t = 3.317$, $p = .008$). Therefore, the model supported the observed pattern that HL students produced tones with higher match rates, and both groups had higher match rates in Grade 3 than in Grade 1. However, the model did not support the observed trend of a lower match of L2 in Grade 5, as there was no significant interaction between group and grade.

The model indicated a target tone effect, with Tone3's match rate lower than the other tones (*p*s < .05). Moreover, group and target interacted: Although L2 students produced all four tones with lower match rates than HL students, the differences were larger in Tone1 ($MD = 0.249$, $z = 5.894$, $p < .001$) and Tone4 ($MD = 0.347$, $z = 7.004$, $p < .001$) and smaller in Tone2 ($MD = 0.103$, $z = 2.954$, $p = .013$) and Tone3 ($MD = 0.192$, $z = 3.254$, $p = .005$). These are in line with the observations that Tone3 was challenging for both groups, and Tone2 was an easier target for L2 students to achieve a similar match rate to

Figure 4. Logistic mixed-model–fitted probabilities of tone match with the significant main effects of group, grade, target, and the significant interaction effects of Group × Target and Grade × Target. Teachers' match rate was plotted in gray as a reference when appropriate. Error bars mark 95% confidence intervals. HL = heritage language; L2 = second language.



their HL peers. Meanwhile, grade and target interacted: Tone2 was the only target that had continually higher average match rates in higher grades and showed a marginally significant difference between Grade 5 and Grade 1 (MD = 0.138, z = 3.000, p = .091). On the contrary, the other tones had the highest average match rates in Grade 3, although differences between grades were not significant (ps > .1).

In summary, the difference in match rates related to backgrounds was not leveled out through bilingual education in the grade levels observed in this study. On the contrary, the HL group produced tones with a higher match than their L2 counterparts. Students achieved a higher match rate in Grade 3, but their tone match rate was not higher in Grade 5. Meanwhile, match rates differed across targets, Tone3 being the most challenging across groups and grades.

Mismatch Patterns

In addition to tone match rates, we were interested in tone mismatch patterns produced by students from different backgrounds. These can provide insights into the intralanguage factors (e.g., which tone pairs are confusable) and interlanguage factors (e.g., whether the patterns are influenced by English). Therefore, confusion matrices were generated for both groups (see Table 2). Only mismatched productions were included, and the cell was bolded if the percentage of this pattern was > 40% among all patterns of this target. Notice that a higher percentage did not indicate more mismatches, since the target might have few mismatches in total (e.g., Tone2➔Tone3 pattern was prevalent in both groups, but the numbers were small).

Both groups produced Tone2➔Tone3 mismatches. Results suggest bidirectional confusion among L2 students, since Tone3➔Tone2 was also dominant in the L2 group's Tone3 patterns, as opposed to the HL group. A GAMM model was used to model the $F0$ contours of Tone2–Tone3 confusion and verify the transcribers' judgment (see Figure 5). The $F0$ contours were similar based on transcribed tones despite the intended targets. If anything, transcribers were lenient when recognizing mismatches: A Tone3 target had to show a strong rising trend to be recognized as Tone2, and a Tone2 target had to show a strong dipping contour to be recognized as Tone3. Unlike the Tone2–Tone3 confusion, which can be explained by phonetic similarity, mismatch patterns in the L2 group included Tone1➔Tone3 (high-level to low-dipping tone) and Tone2➔Tone4 (rising to falling tone).

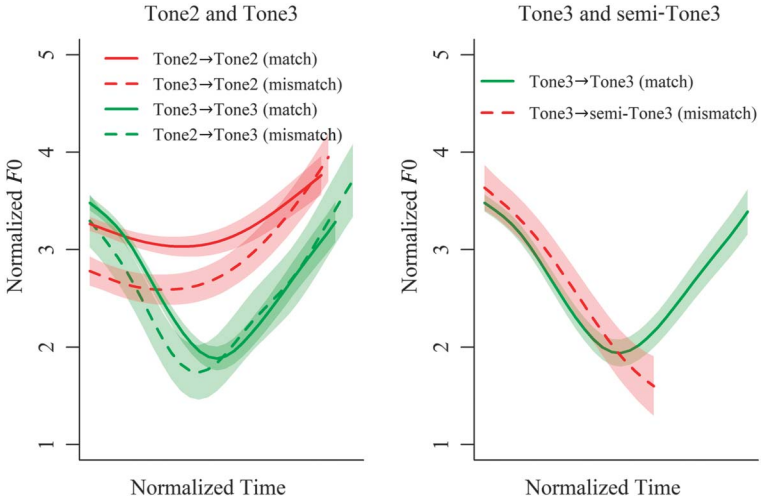
On the other hand, HL students made more uncategorizable productions and produced Tone3 as its semi-sandhi more frequently. All students seemed to be aware of the necessity of tones in monosyllabic words, so the uncategorizable code did not indicate "no tone." Instead, it was used when the production had perceivable differences from the target but could not be assigned to a different category or when the production could be assigned to more than

Table 2. Confusion matrices of heritage language (HL) and second language (L2) groups' production of citation tones.

	Transcribed production					
Target tone	**Tone1**	**Tone2**	**Tone3**	**Tone4**	**Semi-Tone3**	**Uncategorizable**
HL group						
Tone1	—	**10 45.5%**	0 0.0%	1 4.5%	0 0.0%	**11 50.0%**
Tone2	1 4.0%	—	**13 52.0%**	0 0.0%	3 12.0%	8 32.0%
Tone3	9 8.5%	12 11.3%	—	13 12.3%	36 34.0%	36 34.0%
Tone4	4 17.4%	2 8.7%	2 8.7%	—	1 4.3%	**14 60.9%**
L2 group						
Tone1	—	24 38.7%	7 11.3%	10 16.1%	0 0.0%	21 33.9%
Tone2	4 16%	—	**14 56.0%**	1 4.0%	2 8.0%	4 16.0%
Tone3	13 9.6%	**71 52.2%**	—	7 5.1%	11 8.1%	34 25.0%
Tone4	17 23.6%	**29 40.3%**	7 9.7%	—	0 0.0%	19 26.4%

Note. Each row presents a target tone, and each column presents the transcribed production. Transcribed productions include the four citation tones, the semi-sandhi of Tone3 (semi-Tone3), and uncategorizable productions. Each cell contains the number of occurrences and the percentage of this pattern among all patterns of this target (across the row). Prevalent patterns (> 40%) are bolded.

Figure 5. GAMM-fitted *F*0 contours of productions related to Tone2–Tone3 confusion and Tone3–semi-Tone3 confusion. Matched productions are coded in solid lines, and mismatched productions are coded in dashed lines. Ribbons describe 95% confidence intervals. GAMM = generalized additive mixed model; *F*0 = fundamental frequency.



one tone category ambiguously. Therefore, the uncategorizable productions might indicate a more mature development than tone substitution, for the productions were not misrecognized phonemically but needed phonetic refinements (Wong, 2012; Xu et al., 2018). The Tone3➔semi-Tone3 pattern is worth more discussion. Tone3 is realized as [214] in citation tones and as [21] in multisyllabic productions (Duanmu, 2007). To validate this allophonic distinction, we asked 12 linguistically naive Mandarin-L1 listeners to listen to children's Tone3➔semi-Tone3 productions. We found that a monosyllabic semi-Tone3 production out of context was often perceived as Tone4 (six out of 12 listeners) or uncategorizable (four out of 12 listeners). Listeners commented the productions sounded "like a half tone," which is supported by the *F*0 contours in Figure 5. Therefore, we chose to transcribe [21] productions as semi-Tone3 and analyze them as "mismatches" to preserve the perceivable phonetic differences. However, although HL students applied the sandhi rule incorrectly, such a pattern suggested that they were exposed to multisyllabic speech materials and were able to produce the distinctive feature of "low pitch" for Tone3, which probably indicated an intermediate level of tone learning (Wong, 2012; C. Yang, 2016). Note that semi -Tone3 [21] has been observed in monosyllabic Tone3 productions in Taiwanese Mandarin (Duanmu, 2007), although this was coded as mismatch in the present study. In our sample, semi-Tone3 production was not found to have a particular association with a Taiwanese background (all four students and one teacher with Taiwanese background did not produce any Tone3➔semi-Tone3). However, we acknowledge that Tone3 has been undergoing changes in certain varieties of Mandarin and that the effects of incidental exposure to such varieties could not be ruled out.

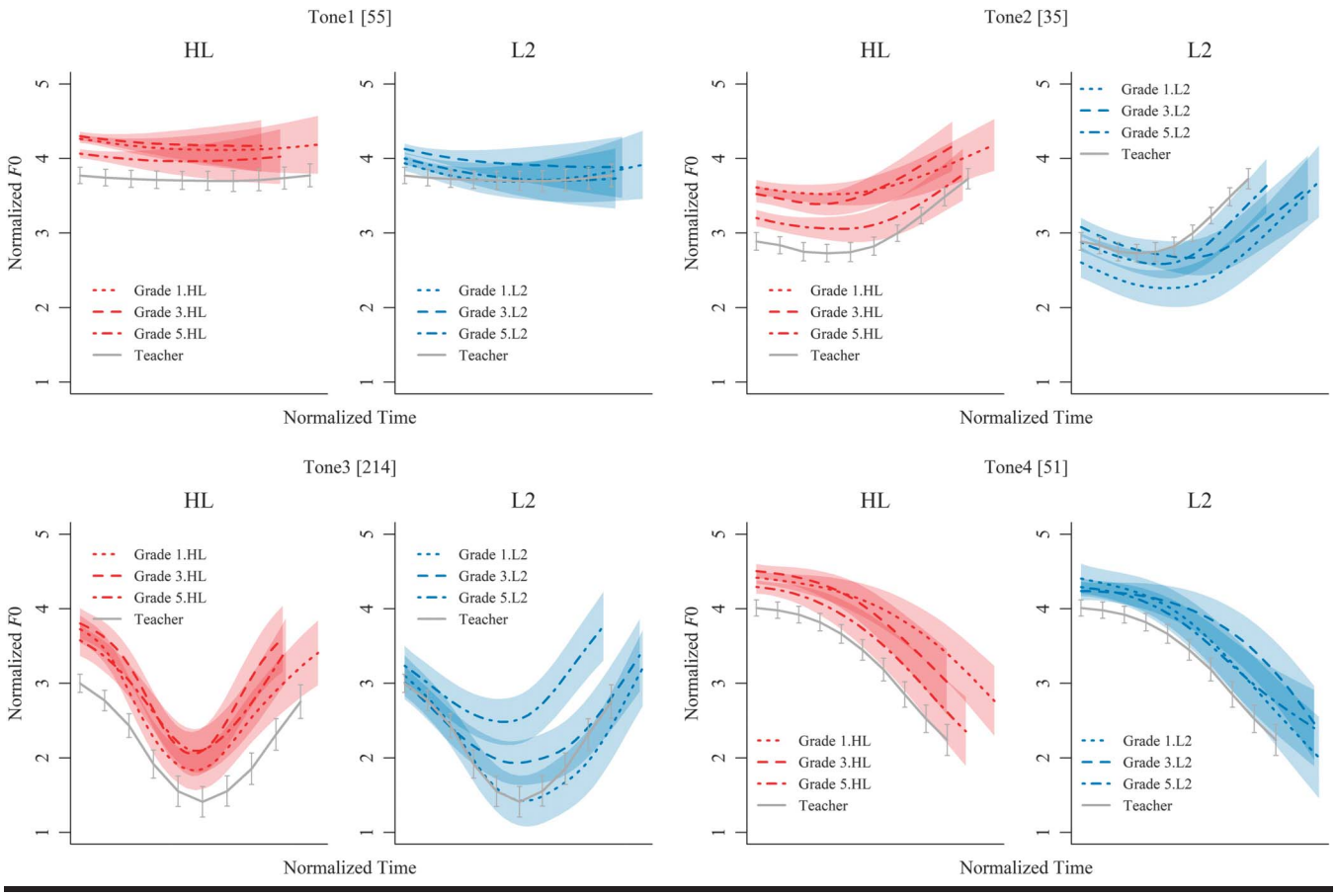
In summary, transcribers' perceptual judgments were validated by acoustic analyses. The L2 group showed a bidirectional confusion of Tone2–Tone3. The HL group exhibited mismatches that phonetically deviated from targets but showed signs of phonemic knowledge, whereas the L2 group produced mismatches that violated distinctive features of the categories.

F0 Contour

Acoustic examinations of matched productions (perceived as accurate) can help understand how bilingual students demonstrate the phonetic specifications of tones. Therefore, *F*0 contours of matched productions were modeled with GAMMs. The model used normalized *F*0 values as the dependent variable and time as the independent variable, with the fixed effects of group, grade, and tone target, as well as the random effects of target word and speaker. Interaction effects were implemented using indexed coding. The compareML function was used to compare between models, and AR(1) models were built to control autocorrelation effects (van Rij et al., 2022). The selected optimal model included parametric and smooth terms of Group × Grade × Tone interaction (adjusted R^2 = .548; see Supplemental Material S3 for model details). This suggested that *F*0 contours were influenced by all three factors. A separate model was built for each tone to present group and grade effects, with teachers' raw data plotted in gray (see Figure 6).

According to Figure 6, most of the important tonal features were produced even in Grade 1. Students' productions resembled those of teachers in higher grades. For example, HL students produced a steeper rising contour for

Figure 6. GAMM-fitted *F*0 contours of each tone target by group and grade with teachers' contours plotted in gray as references. Ribbons and error bars mark 95% confidence intervals. GAMM = generalized additive mixed model; HL = heritage language; L2 = second language; *F*0 = fundamental frequency.



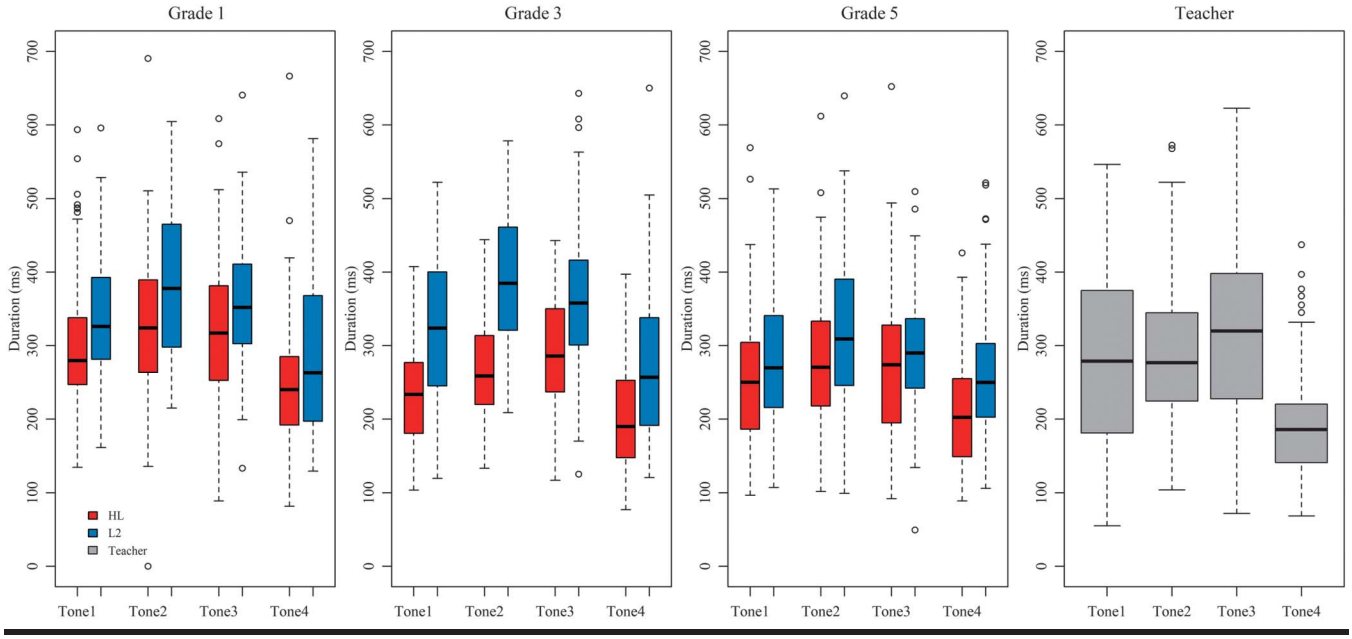
Tone2 and a steeper falling contour for Tone4 in Grade 5. However, L2 students' Tone3 productions did not seem to follow this progression. Instead, in higher grades, L2 students produced Tone3 with higher *F*0 and a shallower dip, thus exhibiting a stronger rising trend: Difference plots of predicted contours using the plot_diff function indicated that in Tone3 productions, Grade 3's *F*0 values were significantly higher than those of Grade 1 from 40% to 60% of the duration (i.e., differences occurred in the middle of the contours), and Grade 5's *F*0 was higher than that of Grade 3 from 18% to 100% of the duration (i.e., differences occurred in most part of the contours). Therefore, although transcribed as matches, L2 students' Tone3 was continually more subjected to being misrecognized as Tone2 in higher grades.

Duration

Duration is an important secondary cue for tone perception (Blicher et al., 1990). Although durations may vary across and within individuals depending on speech rates, the patterns we detected from the data were convincing and can provide insights into students' learning of secondary cues of tones. A plot of raw data by target and group in each grade level is presented in Figure 7, with teachers' data plotted in gray as a reference. It seems that the teachers' duration pattern was comparable with that in L1 literature: Tone3 had the longest duration, and Tone4 had the shortest (Xu, 1997). However, only HL students in Grade 3 and Grade 5 produced a similar pattern, whereas L2 consistently produced Tone2 with a similar duration to Tone3, if not longer.

A linear mixed model was built to verify such observations. The target was duration. Fixed effects included group, grade, and tone target. Random effects included target word and speaker. Models were compared based on AIC. Between-group comparisons were achieved using emmeans, and the *MD*s reported below are in milliseconds. The selected optimal model (see Supplemental Material S4 for model details) suggested a significant group effect: HL students produced tones with significantly shorter durations (MD = 45.981, t = 4.664, p <

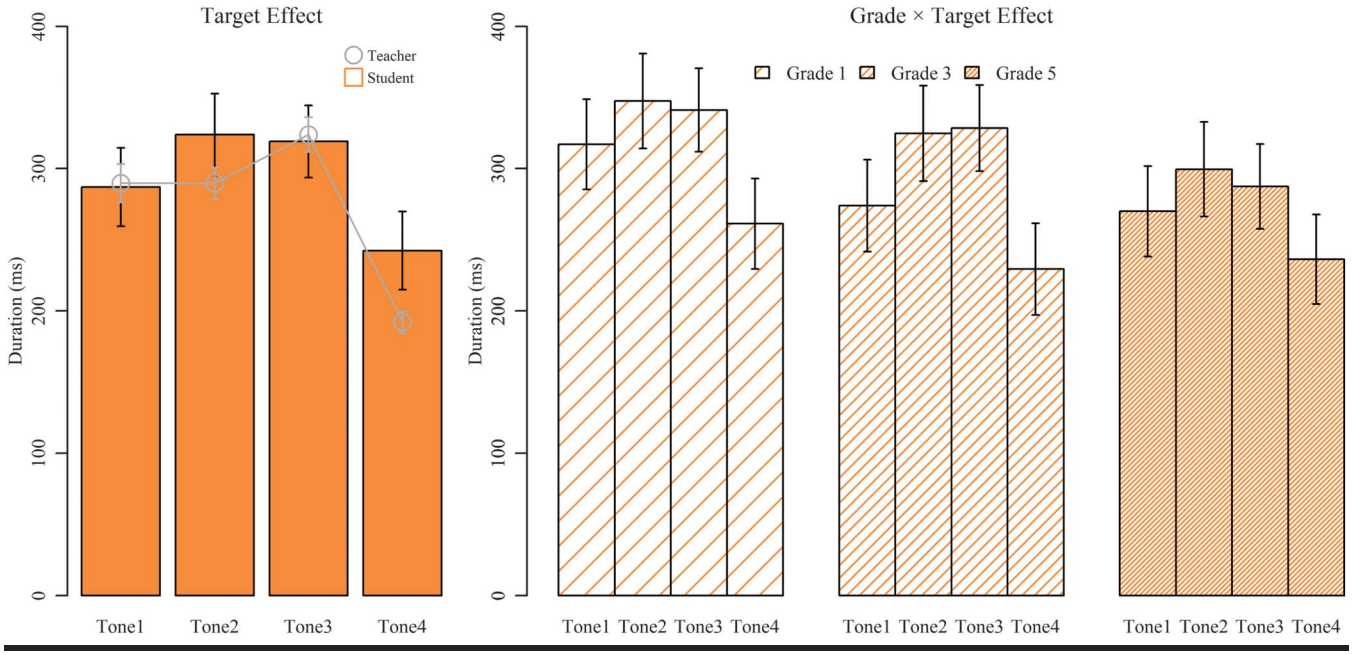
Figure 7. Raw data of the duration of each target by group in each grade level, with teachers' productions plotted in gray as a reference. The boxes and whiskers mark quantiles, and the scattered points mark outliers. HL = heritage language; L2 = second language.



.001). Meanwhile, the model suggested a significant grade effect: Grade 5 students produced a shorter duration than Grade 1 students (MD = 43.405, t = 3.475, p = .007). Such trends are unsurprising as temporal characteristics are related to proficiency in the language (Trofimovich & Baker, 2006).

More relevant to the current study, bilingual students' durations differed across targets (see Figure 8). First, the tone target had a significant effect: Tone4 was significantly shorter than Tone2 (MD = 81.555, t = 4.434, p = .004) and Tone3 (MD = 76.700, t = 4.476, p = .004) but not shorter than Tone1 (MD = 44.662, t = 2.498, p =

Figure 8. Linear mixed-model–fitted duration with significant tone target effect with teachers' productions plotted in gray as a reference and significant Target × Grade effect. Error bars mark 95% confidence intervals.



.491). Tone1, Tone2, and Tone3 were not significantly different (ps > 0.1). That is to say, the model indicated a duration order of Tone3 = Tone2 = Tone1 ≥ Tone4, which does not match the patterns in L1 speakers; that is, Tone3 is usually the longest (Xu, 1997). Such a pattern among bilingual students left Tone2 and Tone3 prone to confusion in the time domain. Moreover, there was a significant interaction between tone and grade. Tone3 was marginally longer than Tone1 in Grade 3 (MD = 54.447, t = 2.957, p = .078), but it was not longer than Tone1 in Grade 5 (MD = 17.522, t = 0.958, p > .999). Therefore, using Tone1's duration as a reference, bilingual students produced Tone3 with a relatively long duration in Grade 3, but this feature was not maintained in Grade 5.

In summary, bilingual students did not pick up the duration difference between Tone2 and Tone3 in the observed time window between Grade 1 and Grade 5. The long-duration feature of Tone3 seemed reduced in Grade 5. Moreover, L2 students confused these two tones in the time domain, as they produced Tone2 with a longer duration on average.

Discussion

The current study presented evidence on how elementary students in a Mandarin–English bilingual program in Western Canada produced Mandarin citation tones and how various factors internal and external to the two languages were associated with their tone productions. Both transcription-based and acoustic analyses were included to examine the match rates, mismatch patterns, and acoustic realizations of tone productions. Match analyses suggest that HL students had advantages over their L2 peers, and the latter did not catch up in higher grades. Students achieved higher match rates in Grade 3 than Grade 1 but not higher in Grade 5. Moreover, both groups achieved high average match rates in Tone2 but had low match rates in Tone3. In terms of mismatches, those in the HL group tended to partly demonstrate the features of the targets (e.g., Tone3➔semi-Tone3), but those in the L2 group violated the target features (e.g., Tone4➔Tone2). Meanwhile, both groups produced Tone2–Tone3 confusion, but L2 more frequently simplified Tone3 to Tone2. In terms of acoustic measurements, both groups produced $F0$ contours that resembled teacher and L1 speakers, but L2 students did not emphasize the low-dipping contour of Tone3 in higher grades. Neither group of students fully acquired the duration features of Mandarin tones up to Grade 5 compared to teachers' and L1 speakers' patterns. The next sections discuss the results in relation to factors of bilingual pronunciation learning and present implications for researchers, theorists, and educators.

Impacts of Intra-, Inter-, and Extralanguage Factors

The first research question is to address intralanguage factors and identify specific tone target(s) that is the most challenging for bilingual students to acquire and produce. Results support the hypothesis that Tone3 is the most challenging tone target. Tone3 has varied phonological realizations (Xu et al., 2018), and advanced motor skills are required to produce its low pitch and compound $F0$ contour (Wong, 2012). Therefore, Tone3 is the latest-developmental citation tone in Mandarin and is challenging for L2 learners (Wang et al., 2003; Wong, 2012). In the current study, Tone3 had the lowest match rates for both HL and L2 students. Meanwhile, Tone3 involves complex acoustic characteristics, such as a low, dipping $F0$ contour, a creaky voice that often co-occurs with the low pitch (Kuang, 2017), and the longest duration (Blicher et al., 1990; Xu, 1997). Monolingual children spend years refining these phonetic specifications (Rhee et al., 2020; Wong, 2012). These cues also appeared challenging for bilingual students: Neither group produced Tone3 with a longer duration than Tone2 at any of the three grade levels we observed. In this sense, bilingual children's tone production depends on the complexity of the tone targets and shows consistency across language background groups.

The second research question addresses interlanguage factors and aims to identify cross-linguistic influences from English. According to PAM-S, since English does not assign as much linguistic significance to pitch information, bilingual learners, especially L2 students, may have difficulty perceiving and contrasting Mandarin tone categories. They may assimilate Mandarin tones into other prosodic categories in English (e.g., Tone2 [35] as question and Tone4 [51] as statement intonations). They may also produce more categorical substitutions, especially between the categories that are phonetically similar (e.g., Tone2 [35]–Tone3 [214]; So & Best, 2010). Results partly support PAM-S's theoretical accounts. Tone2 was the only tone where L2 and HL students reached similar levels of high match rates. According to teachers' reflections, they used the question intonation in English to facilitate the learning of Tone2 (Lin et al., 2024). With this strategy, L2 students acquired the category of Tone2 efficiently, which is in line with Wang et al.'s (2003) study where adult L2 learners showed the most improvement in Tone2 after a short period of training. Moreover, L2 students produced Tone3➔Tone2 mismatches frequently. This pattern was found prevalent among English-L1 learners (Wang et al., 2003) and also reported among younger Mandarin monolingual children (Wong, 2012; Zhu, 2002). Thus, this pattern may be related to the

difficulty distinguishing the phonemic categories due to the lack of such a contrast in English, but it could also be explained by intrinsic difficulty of Tone3 and the simplification of the phonetic target (falling–rising➔rising). Furthermore, L2 students produced tone substitutions that violated the target features (e.g., Tone4➔Tone2, even though they could be assimilated into two distinctive prosodic categories in English). Such patterns were not highlighted in PAM-S. They may be better explained by limited linguistic significance assigned to tone categories due to English influences (So & Best, 2010) and a generally immature phonetic representation of tone categories (Flege & Bohn, 2021).

Even for the HL students, their tone production skills were not like younger monolingual counterparts in previous research despite long-term and frequent exposure to Mandarin at home and at school, which suggests the influence of English on bilingual tone learning. Zhu (2002) reported no tone error in citation tones produced by monolingual children who were 2–4.5 years of age, while the HL students produced tone mismatch at both the phonetic and phonemic levels. In addition to the patterns reviewed above that were outstanding in L2 students' productions, both groups produced Tone1➔Tone2 and Tone2➔Tone3 substitutions. The difficulty in Tone1 was likely related to the influence of the English phonology. Although a level *nucleus tone* (intonation pattern) exists in English to express continuation (Levis & Wichmann, 2015), it is not produced as consistently as using a rising tone for questions (Hudson et al., 2019). Therefore, English bilingual children are likely to take longer to establish an independent category for Tone1 and assimilate it to Tone2 (Flege & Bohn, 2021; So & Best, 2010). As for Tone2➔Tone3, this pattern was not commonly found in monolingual children (Wong, 2012; Zhu, 2002) but was reported among adult English-L1 learners of Mandarin: Hao (2012) found a bidirectional confusion between Tone2 and Tone3 and attributed it to the intrinsic difficulty of this tone pair. The present study offers an alternative explanation related to English influences and pronunciation instruction. As the Chinese teachers used questions in English to facilitate the learning of Tone2, they often demonstrated the question intonation with a long duration and exaggerated pitch contour (Lin et al., 2024). As a side effect, students in the present study produced Tone2 with a lower *F*0 onset (see Figure 6; L2 students) and a similar duration to Tone3 (see Figure 7; both groups). This is plausible for English intonation production, since the nucleus tone often involves a low-rising contour, and the low feature can be extended (Hedberg et al., 2014). However, the onset *F*0 and the timing of *F*0 valley are important cues for Mandarin listeners to distinguish Tone2 from Tone3 (Wang & Li, 2010; Wong, 2012). Therefore, Tone2➔Tone3 may be a result of bilingual children using the question intonation in English to approximate Tone2.

The third research question addresses extralanguage factors such as Mandarin input in learners' environments (Flege & Bohn, 2021), which were operationalized as home language backgrounds (early and ongoing home input) and grade levels in the Mandarin program (recent school input). It remained unknown how these factors interact in school-aged children in the context of bilingual education. Existing evidence had shown the effectiveness of two-way bilingual education in pronunciation teaching and learning, where the differences related to home language background can be leveled out through immersive input at school (Menke, 2017; Nance, 2021). Accordingly, it was hypothesized that strong home input and increased school input would be positively associated with tone learning outcomes and that the L2 learners would catch up to their HL peers in higher grades. The results, however, do not provide straightforward support for the hypotheses. L2 students did not catch up to their HL peers in higher grades and showed increased phonetic deviation in their Tone3 production in higher grades, and even for HL students, their tone match rates plateaued between Grades 3 and 5 did not reach the level like their teachers. This indicated that bilingual students' tone learning of Mandarin as a minority language was affected by the limited speech input and output in the English-dominant society and was more protracted compared to monolingual children (Wong, 2012). Notably, parents reported that larger proportions of students were English-dominant in Grades 3 and 5 than in Grade 1 (see Table 1). Meanwhile, in higher grades, the pedagogical focus shifted from pronunciation and basic vocabulary to reading and writing, which could also have altered the speech input and output at school (Alberta Education, 2006; Lin et al., 2024). A possible alternative explanation is that higher grade students, with more advanced lexical knowledge in Mandarin, were willing to produce target words spontaneously, even when the encoding of tone information was inaccurate or unstable. However, in the present study, students were not noted to produce the target words more spontaneously in Grade 5 than Grade 3. Thus, the alternative explanation cannot be supported by the current data and requires further investigation.

SLM-r (Flege & Bohn, 2021) did not make specific hypotheses about how home input and school input would have different impacts on speech learning, but the results of the current study suggested that the effects are not simply additive and were influenced by the complexity of targets. The pitch and duration features of Tone3 were reduced for the L2 group in higher grades, which means their developmental trajectories were not parallel to HL peers and were probably more subjected to attrition. Such

developmental trajectories may be related to L2 students' limited home input in Mandarin at early ages. This is in line with evidence that infants in nontonal language environments learn to ignore suprasegmental information in word recognition (Singh et al., 2008). Such evidence did not support Wong's (2012) hypothesis that the protracted refinement of tones was related to immature physiology. The loss of phonetic features in higher grades among L2 students suggested that tone learning is more related to integrating phonetic cues through perceptual learning and social interactions (Kuhl et al., 2003; Rhee et al., 2020).

Implications and Future Directions

In addition to providing the evidence aforementioned, the current study has several implications for research methods, bilingual speech theories, and pedagogical practices. In terms of research methods, the current study did not use overall tone match (e.g., percent tone correct; Shriberg et al., 1997) as a single psychometric score. Instead, we examined the match of each production as a function of tone target and controlled the effects of speaker and target word using mixed modeling. It is not uncommon to observe specific targets separately in language and speech development, as each category and even item may have different levels of complexity (McMillen et al., 2022; Zhu, 2002). The current study again proved the merit of such analyses, since Tone3 was significantly more challenging across groups and grade levels. In addition, the current study investigated both phonemic categories and phonetic characteristics of tones, which revealed more details in the learning process. Instead of forcing transcribers to choose a phonemic category (e.g., Zhu, 2002), the current study allowed transcribers to indicate that a tone production was an inappropriate allophonic realization or was uncategorizable (Wang et al., 2003). With such allophonic considerations, students' match rates did not reach teachers' level after years of school learning. This is different from the evidence in monolingual and bilingual children that tones are mastered early (Holm & Dodd, 2006; Mok & Lee, 2018; Zhu, 2002). Meanwhile, it needs to be acknowledged that tone mismatch may signal breakdowns at the lexical, phonological, or phonetic levels, which could not be fully disentangled with the present data, but for early learners, the connections between phonetic forms, phonemic forms, and lexical forms are all important components of speech learning (van Leussen & Escudero, 2015; Werker & Curtin, 2005). In future research, especially research among more advanced learners, a variety of tasks in addition to picture naming can be used to measure different components of tone production and perception skills (e.g., identification, mimicry, reading; Hao, 2012), and the role of specific lexical items can be further investigated (McMillen et al., 2022). Indeed, both groups reached a functional level of match, but the confusion matrices revealed that the two groups had different tone patterns. Meanwhile, even among matched productions, acoustic analyses suggested that bilingual students might not utilize certain phonetic features in their tone productions (i.e., duration). This supports Wong's (2012) argument that phonemic transcriptions may overestimate children's tonal skills, and children's tone development would appear more protracted when phonetic details are considered. It is noteworthy, however, that acoustic analyses should be linked to listener perception to provide functional implications (Munro & Derwing, 2020). Therefore, future research will continue to include other acoustic measurements such as phonation and contour parameters such as *F*0 shift and *F*0 range (Kuang, 2017; Wong, 2012) and relate them to perceptual judgments such as intelligibility and accentedness (Munro & Derwing, 2020).

In terms of theoretical implications, the current study supported PAM-S and SLM-r. Bilingual students' tone productions were related to English transfer (So & Best, 2010) and Mandarin speech input (Flege & Bohn, 2021). Results also expanded these theories in the suprasegmental domain. First, the results emphasized the effects of universal challenges in speech production (Wong, 2012) by showcasing the difficulty of producing Tone3 across groups. Second, the current study suggests the unique value of early home input. At least in the context of learning a tonal language in a nontonal environment, HL students who received Mandarin home input were able to achieve high match rates and maintain important phonetic features in their production, whereas their L2 peers seemed to be less sensitive to categorical differences and phonetic details despite their current exposure to Mandarin at school. Thus, it seems that "earlier is better." However, HL students' advantage might also be related to the high-quality and continual input at home and in the community (Flege & Bohn, 2021). Meanwhile, HL students also seemed to have plateaued in their tone production skills in higher grades, possibly related to limited exposure to Mandarin at home as they shifted their language dominance, and limited pronunciation instruction at school due to the curriculum design. Therefore, our results did not suggest a decisive effect of early exposure (Bedore et al., 2016). Instead, the results suggested the advantage of early exposure among heritage speakers needs to be supported by continual exposure to the HL (Chang et al., 2011; J. Yang & Fox, 2017). "Input," an important factor of L2 speech learning (Flege & Bohn, 2021), has often been measured through length of residence or the concurrent frequency of L1–L2 use (Flege, 2021). With the current results, we advocate for a more fine-grained and comprehensive operational definition of "input" that encompasses the quantity and quality of bilingual students' language

experiences across different settings, including language use and language activities at home, in school, and in the community, as well as the nature of language instruction. Future research should continue to examine how quantity and quality of input, measured numerically, explain variance in speech outcomes of children with diverse bilingual experiences. In addition to cross-sectional studies, longitudinal, experimental studies should be conducted to understand the effects of improved quantity and quality of input, especially the speech input that involves meaningful social interactions.

Finally, the current study provides implications for pedagogical practices. First, it provides evidence on pronunciation learning, which is understudied in the context of bilingual education. Results showed that both HL and L2 students were able to produce citation tones with core phonetic features as early as in Grade 1 and achieve match rates well above chance levels. In addition, both groups produced tones with increased match rates in Grade 3. Furthermore, the HL group maintained an advantage in tone production and did not seem to experience significant attrition in Mandarin tone production, as opposed to Mandarin-HL children with low Mandarin exposure whose speech production presented with great influences from the L2 (J. Yang & Fox, 2017). These findings highlighted the effectiveness of two-way bilingual education design in fulfilling its functions of L2 learning and HL maintenance. On the other hand, the differences related to home language backgrounds were not leveled out through bilingual education at the elementary level, which provides different evidence compared to studies in bilingual programs of two Indo-European languages (cf. Menke, 2017; Nance, 2021, on other bilingual programs). On the contrary, L2 students had difficulties developing pronunciation skills continually, especially in terms of further refining the phonetic specifications of tones. Furthermore, a plateauing effect of tone productions was observed in both groups of students, similar to the phonetic fossilization observed in one-way French immersion education (Netelenbos et al., 2016), despite the expected merit of two-way bilingual design (Cummins, 1979; Menke, 2017; Nance, 2021). In summary, bilingual students did not receive enough input to achieve ideal learning outcomes. Such a result can be attributed to typological differences between Mandarin and English (compared to the Indo-European language pairs in previous evidence) and generally limited Mandarin input in the community (compared to the availability of Spanish input in certain areas in the United States), and it also seems to be related to the curriculum design. Since Mandarin pronunciation was mainly addressed through the instruction on sound–letter relationships in lower grades (Alberta Education, 2006), teachers were hesitant to teach pronunciation in higher grades (Lin et al., 2024). In addition, it is common for early-arrival children to shift their language dominance to English within the first year of immigration (Jia & Aaronson, 2003). Although HL students maintained an advantage compared to L2 peers, most of them preferred using English due to the English-dominant environment at school and in the community (Lin et al., 2024). Therefore, policymakers and educators may consider making pronunciation a long-term goal and support the development of the societal minority language to prevent HL/L2 attrition among students with diverse backgrounds. Second, the current study can help identify practical strategies for Mandarin pronunciation instruction. All measurements clearly indicated that Tone3 was challenging and was often confused with Tone2. The present study directs educators' attention to phonetic details in pronunciation instructions in addition to the establishment of phonemic categories. Educators may benefit from professional development programs to understand the phonetic features of these tone targets. For example, educators can emphasize that Tone3 is not only dipping but also low, often accompanied by a creaky voice (Kuang, 2017). They can also stress that Tone3 is produced with a longer duration, which allows it to be distinguished from Tone2 (Blicher et al., 1990) and gives students more time to articulate its complex *F*0 contour (Wong, 2012).

The current study is among the first to document school-aged children's learning of tones in a nontonal language environment, investigate the pronunciation learning outcomes of students enrolled in bilingual education programs, and use both phonemic and phonetic analyses to examine such learning. It draws attention to the multifaceted factors that influence bilingual speech development and calls for researchers and educators to continually support bilingual children's pronunciation learning in a minority-language context.

Data Availability Statement

Data are available from the corresponding author upon request. Some intermediate data (mean duration of tones produced by each speaker group) are directly available in the supplemental materials.

Acknowledgments

This work was supported by the Social Sciences and Humanities Research Council (SSHRC) under the Insight Grant (435-2017-1086) to Fangfang Li and Karen Pollock and the Vanier Canada Graduate Scholarship through SSHRC (CGV—163274) to Youran Lin. The authors

would also like to thank the participating schools, principals, teachers, children, and parents; Edmonton Chinese Bilingual Education Association for their continued dedication to bilingual education and research; Xiaozhu Chen, Lujia Yang, Nan Xing, and Minjia Tao for their help with data collection and transcription; and Benjamin V. Tucker and Yvan Rose for suggestions on the methodology.

References

Alberta Education. (2006). *Chinese language arts kindergarten to Grade 9: International languages, programs of study*. https://education.alberta.ca/international-languages-k-6/programs-of-study/

Baker, W., & Trofimovich, P. (2005). Interaction of native- and second-language vowel system(s) in early and late bilinguals. *Language and Speech, 48*(1), 1–27. https://doi.org/10.1177/00238309050480010101

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bedore, L. M., Peña, E. D., Griffin, Z., & Hixon, G. (2016). Effects of age of English exposure, current input/output, and grade on bilingual language performance. *Journal of Child Language, 43*(3), 687–706. https://doi.org/10.1017/s0305000915000811

Best, C. T., & Tyler, M. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (pp. 13–34). John Benjamins. https://doi.org/10.1075/lllt.17.07bes

Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics, 18*(1), 37–49. https://doi.org/10.1016/s0095-4470(19)30357-2

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* (Version 6.3.14) [Computer software]. http://www.praat.org/

Canadian Charter of Rights and Freedoms, s 23, Part I of the Constitution Act of 1982, being Schedule B to the Canada Act 1982 (UK), c11 (1982).

Chang, C. B., Yao, Y., Haynes, E. F., & Rhodes, R. (2011). Production of phonetic and phonological contrast by heritage speakers of Mandarin. *The Journal of the Acoustical Society of America, 129*(6), 3964–3980. https://doi.org/10.5070/p75p6693q0

Chao, Y. R. (1930). A system of tone letters. *Le Maître Phonétique, 8*(45), 24–27.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*(2), 222–251. https://doi.org/10.2307/1169960

Dicks, J., & Genesee, F. (2017). Bilingual education in Canada. In O. García, A. M. Y. Lin, & S. May(Eds.), *Bilingual and multilingual education* (3rd ed., pp. 453–467). Springer. https://doi.org/10.1007/978-3-319-02258-1_32

Duanmu, S. (2007). *The phonology of standard Chinese* (2nd ed.). OUP Oxford. https://doi.org/10.1093/oso/9780199215782.001.0001

Flege, J. E. (2021). New methods for second language (L2) speech research. In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 119–156). Cambridge University Press. https://doi.org/10.1017/9781108886901.004

Flege, J. E., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Goldstein, B. A., Fabiano, L., & Washington, P. S. (2005). Phonological skills in predominantly English-speaking, predominantly Spanish-speaking, and Spanish–English bilingual children. *Language, Speech, and Hearing Services in Schools, 36*(3), 201–218. https://doi.org/10.1044/0161-1461(2005/021)

Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics, 32*(3), 395–421. https://doi.org/10.1016/s0095-4470(03)00016-0

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. https://doi.org/10.1016/j.wocn.2011.11.001

Harada, T. (2007). The production of voice onset time (VOT) by English-speaking children in a Japanese immersion program. *International Review of Applied Linguistics in Language Teaching, 45*(4), 353–378. https://doi.org/10.1515/iral.2007.015

Hedberg, N., Sosa, J. M., & Görgülü, E. (2014). The meaning of intonation in yes–no questions in American English: A corpus study. *Corpus Linguistics and Linguistic Theory, 13*(2), 321–368. https://doi.org/10.1515/cllt-2014-0020

Hedlund, G., & Rose, Y. (2020). *Phon 3.1* [Computer software]. https://phon.ca

Holm, A., & Dodd, B. (2006). Phonological development and disorder of bilingual children acquiring Cantonese and English. In H. Zhu & B. Dodd (Eds.), *Phonological development and disorders in children: A multilingual perspective* (pp. 286–325). De Gruyter. https://doi.org/10.21832/9781853598906-014

Hudson, T., Setter, J., & Mok, P. (2019). Nuclear tones in Hong Kong and British English. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 320–323). Australasian Speech Science and Technology Association. https://centaur.reading.ac.uk/80883/

Jia, G., & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics, 24*(1), 131–161. https://doi.org/10.1017/S0142716403000079

Kan, R. T., & Schmid, M. S. (2019). Development of tonal discrimination in young heritage speakers of Cantonese. *Journal of Phonetics, 73,* 40–54. https://doi.org/10.1016/j.wocn.2018.12.004

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America, 142*(3), 1693–1706. https://doi.org/10.1121/1.5003649

Kuhl, P. K., Tsao, F. M., & Liu, H. M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences, 100*(15), 9096–9101. https://doi.org/10.1073/pnas.1532872100

Levis, J. M., & Wichmann, A. (2015). English intonation—Form and meaning. In M. Reed & J. M. Levis (Eds.), *The handbook of English pronunciation* (pp. 139–155). John Wiley & Sons.

Lin, Y., Li, F., & Pollock, K. E. (2024). Pronunciation teaching in minority languages: Perspectives of elementary school teachers in a Chinese–English bilingual program in Canada. *Cogent Education, 11*(1), Article 2356432. https://doi.org/10.1080/2331186X.2024.2356432

Lu, L., & Liu, H. S. (1998). *The Peabody Picture Vocabulary Test–Revised in Chinese*. Psychological Publishing.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Marinova-Todd, S. H., Zhao, J., & Bernhardt, M. (2010). Phonological awareness skills in the two languages of Mandarin–English bilingual children. *Clinical Linguistics & Phonetics, 24*(4–5), 387–400. https://doi.org/10.3109/02699200903532508

McMillen, S., Anaya, J. B., Peña, E. D., Bedore, L. M., & Barquin, E. (2022). That's hard! Item difficulty and word characteristics for bilinguals with and without developmental language disorder. *International Journal of Bilingual Education and Bilingualism, 25*(5), 1838–1856. https://doi.org/10.1080/13670050.2020.1832039

Meckelborg, A., Luu, M., Nguyen, T., Lin, Y., Li, F., & Pollock, K. (2024). Acquisition of Mandarin tones by Canadian first graders: Effect of prior exposure to tonal and non-tonal languages. *The Journal of the Acoustical Society of America, 155*(2), 1608–1623. https://doi.org/10.1121/10.0024985

Menke, M. R. (2015). How native do they sound? An acoustic analysis of the Spanish vowels of elementary Spanish immersion students. *Hispania, 98*(4), 804–824. https://doi.org/10.1353/hpn.2015.0123

Menke, M. R. (2017). Phonological development in two-way bilingual immersion: The case of Spanish vowels. *Journal of Second Language Pronunciation, 3*(1), 80–108. https://doi.org/10.1075/jslp.3.1.04men

Mok, P. P. K., & Lee, A. (2018). The acquisition of lexical tones by Cantonese–English bilingual children. *Journal of Child Language, 45*(6), 1357–1376. https://doi.org/10.1017/s0305000918000260

Munro, M. J., & Derwing, T. M. (2020). Collecting data in L2 pronunciation research. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching Conference* (pp. 2380–9566). Northern Arizona University.

Nance, C. (2021). Scottish Gaelic revitalisation: Progress and aspiration. *Journal of Sociolinguistics, 25*(4), 617–627. https://doi.org/10.1111/josl.12508

Netelenbos, N., Li, F., & Rosen, N. (2016). Stop consonant production of French immersion students in Western Canada: A study of voice onset time. *International Journal of Bilingualism, 20*(3), 346–357. https://doi.org/10.1177/1367006914564566

Paradis, J. (2011). Individual differences in child English second language acquisition: Comparing child-internal and child-external factors. *Linguistic Approaches to Bilingualism, 1*(3), 213–237. https://doi.org/10.1075/lab.1.3.01par

Paradis, J. (2023). Sources of individual differences in the dual language development of heritage bilinguals. *Journal of Child Language, 50*(4), 793–817. https://doi.org/10.1017/S0305000922000708

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rhee, N., Chen, A., & Kuang, J. (2020). Integration of spectral cues in the development of Mandarin tone production. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 3135–3138). Australasian Speech Science and Technology Association.

Shi, F., & Wang, P. (2006). A statistical analysis of tone groups in Beijing Mandarin. *Contemporary Linguistics, 8*(4), 324–333.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The percentage of consonants correct (PCC) metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research, 40*(4), 708–722. https://doi.org/10.1044/jslhr.4004.708

Sieg, S. R., Fabiano, L., & Barlow, J. (2023). Substitution errors and the role of markedness in bilingual phonological acquisition. *Journal of Speech, Language, and Hearing Research, 66*(12), 4699–4715. https://doi.org/10.1044/2023_JSLHR-23-00116

Singh, L., White, K. S., & Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development, 4*(2), 157–178. https://doi.org/10.1080/15475440801922131

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and Speech, 53*(2), 273–293. https://doi.org/10.1177/0023830909357156

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*(01), 1–30. https://doi.org/10.1017/s0272263106060013

van Leussen, J. W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology, 6,* Article 1000. https://doi.org/10.3389/fpsyg.2015.01000

van Rij, J., Wieling, M., Baayen, R., & van Rijn, H. (2022). *itsadug: Interpreting time series and autocorrelated data using GAMMs* (R package Version 2.4.1) [Computer software].

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America, 113*(2), 1033–1043. https://doi.org/10.1121/1.1531176

Wang, Y., & Li, M. (2010). The effects of tone pattern and register in perceptions of Tone 2 and Tone 3 in Mandarin. *Acta Psychologica Sinica, 42*(9), 899–908. https://doi.org/10.3724/SP.J.1041.2010.00899

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1*(2), 197–234. https://doi.org/10.1080/15475441.2005.9684216

Wong, P. (2012). Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic Mandarin lexical tone productions. *Journal of Phonetics, 40*(1), 141–151. https://doi.org/10.1016/j.wocn.2011.10.005

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press. https://doi.org/10.1201/9781315370279

Xu Rattanasone, N., Tang, P., Yuen, I., Gao, L., & Demuth, K. (2018). Five-year-olds' acoustic realization of Mandarin tone sandhi and lexical tones in context are not yet fully adult-like. *Frontiers in Psychology, 9,* Article 817. https://doi.org/10.3389/fpsyg.2018.00817

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics, 25*(1), 61–83. https://doi.org/10.1006/jpho.1996.0034

Xu, Y. (2013). ProsodyPro-A tool for large-scale systematic prosody analysis. *TRASP Aix-en-Provence,* 7–10.

Yang, C. (2016). *The acquisition of L2 Mandarin prosody: From experimental studies to pedagogical practice (Vol. 1)*. John Benjamins. https://doi.org/10.1075/bpa.1

Yang, J., & Fox, R. A. (2017). L1–L2 interactions of vowel systems in young bilingual Mandarin–English children. *Journal of Phonetics, 65,* 60–76. https://doi.org/10.1016/j.wocn.2017.06.002

Yang, J., & Liu, C. (2012). Categorical perception of lexical tone in 6 to 8-year-old monolingual and bilingual children. *International Journal of Asian Language Processing, 22*(2), 49–62.

Yang, L., Lin, Y., Pollock, K. E., & Li, F. (2021). *Accuracy of spontaneous and imitative productions in children learning Mandarin in a bilingual program* [Conference presentation]. The International Child Phonology Conference (ICPC2021), virtual.

Yao, Y., Chan, A., Fung, R., Wu, W. L., Leung, N., Lee, S., & Luo, J. (2020). Cantonese tone production in pre-school Urdu–Cantonese bilingual minority children. *International Journal of Bilingualism, 24*(4), 767–782. https://doi.org/10.1177/1367006919884465

Zhang, J. (2018). A comparison of tone normalization methods for language variation research. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Association for Computational Linguistics. https://aclanthology.org/Y18-1095/

Zhao, J., & Bernhardt, B. M. (2012). *Mandarin single-word elicitation tool for phonology*. Phonological Development Tools and Cross-linguistic Phonology Project. https://phonodevelopment.sites.olt.ubc.ca/mandarin-picture-elicitation_2012/

Zhu, H. (2002). *Phonological development in specific contexts: Studies of Chinese-speaking children*. Multilingual Matters. https://doi.org/10.21832/9781853595899

Appendix

The 32 Citation Tone Targets (Monosyllabic Words), Sorted From the Most Spontaneous to the Least Spontaneous Within Each Tone Category

Tone	Word	Gloss	Pinyin	IPA	Spontaneity (%)
Tone1	三	Three	san1	/san^{55}/	98
Tone1	八	Eight	ba1	/pa^{55}/	98
Tone1	书	Book	shu1	/ʂu^{55}/	81
Tone1	吃	Eat	chi1	/tʂʰʅ^{55}/	79
Tone1	车	Car	che1	/tʂʰɤ^{55}/	79
Tone1	山	Mountain	shan1	/ʂan^{55}/	69
Tone1	灯	Light	deng1	/təŋ^{55}/	53
Tone1	虾	Shrimp	xia1	/ɕja^{55}/	47
Tone2	蓝	Blue	lan2	/lan^{35}/	93
Tone2	鱼	Fish	yu2	/y^{35}/	85
Tone2	球	Ball	qiu2	/tɕʰjo͡ʊ^{35}/	74
Tone2	糖	Candy	tang2	/tʰɑŋ^{35}/	72
Tone2	门	Door	men2	/mən^{35}/	65
Tone2	圆	Circle	yuan2	/ɥɛn^{35}/	64
Tone2	床	Bed	chuang2	/tʂʰwɑŋ^{35}/	62
Tone3	五	Five	wu3	/u^{214}/	99
Tone3	手	Hand	shou3	/ʂo͡ʊ^{214}/	95
Tone3	水	Water	shui3	/ʂwe͡ɪ^{214}/	94
Tone3	狗	Dog	gou3	/ko͡ʊ^{214}/	94
Tone3	马	Horse	ma3	/ma^{214}/	90
Tone3	紫	Purple	zi3	/tsɿ^{214}/	76
Tone3	雨	Rain	yu3	/y^{214}/	73
Tone3	脚	Foot	jiao3	/tɕjɑ͡ʊ^{214}/	65
Tone3	碗	Bowl	wan3	/wan^{214}/	49
Tone4	二	Two	er4	/ɐɚ^{51}/	98
Tone4	绿	Green	lv4 (lü4)	/ly^{51}/	96
Tone4	饭	Rice	fan4	/fan^{51}/	72
Tone4	肉	Meat	rou4	/ʐo͡ʊ^{51}/	71
Tone4	饿	Hungry	e4	/ɤ^{51}/	68
Tone4	热	Hot	re4	/ʐɤ^{51}/	68
Tone4	站	Stand	zhan4	/tʂan^{51}/	62
Tone4	菜	Vegetable	cai4	/tsʰa͡ɪ^{51}/	57

Note. IPA = International Phonetic Alphabet.