ORIGINAL RESEARCH
published: 17 June 2022
doi: 10.3389/fcomm.2022.896013

Production and Perception of Mandarin Laryngeal Contrast: The Role of Post-plosive F0

Yuting Guo[†] *and Harim Kwon**[†]

Linguistics Program, Department of English, George Mason University, Fairfax, VA, United States

This study examines the relation between plosive aspiration and post-plosive f0 (fundamental frequency) in the production and perception of the laryngeal contrast in Mandarin. Production data from 25 Mandarin speakers showed that, in word onsets, VOTs (voice onset time) of aspirated and unaspirated plosives were different, as expected. At the same time, the speakers produced different post-plosive f0 between aspirated and unaspirated plosives, but the difference varied according to the lexical tones – post-aspirated f0 was higher than post-unaspirated f0 in high-initial tones (i.e., lexical tones with high onset f0), but the pattern was the opposite and less robust in low-initial tones. In the perception of the same participants, VOT was the primary cue to aspiration but, when VOT was ambiguous, high post-plosive f0 yielded more aspirated responses in general. We claim that the asymmetry in f0 perturbation between high-initial and low-initial tones in production arises from different laryngeal maneuvers for different tonal targets. In low-initial tones, in which the vocal folds are slack and the glottal opening is wider, aspirated plosives have a lower subglottal air pressure than unaspirated plosives at the voicing onset, resulting in lower post-aspirated f0 than post-unaspirated f0. But in high-initial tones, the vocal folds are tense, which requires a higher trans-glottal pressure threshold to initiate phonation at the onset of voicing. As a result, the subglottal pressure does not decrease as much. Instead, the faster airflow in aspirated than unaspirated plosives gives rise to the pattern that post-aspirated f0 is higher than post-unaspirated f0. Regardless of this variation in production, our perception data suggest that Mandarin listeners generalize the f0 perturbation patterns from high-initial tones and associate high post-plosive f0 with aspirated plosives even in low-initial tone contexts. We cautiously claim that the observed perceptual pattern is consistent with the robustly represented production pattern, as high-initial tones are more prevalent and salient in the language and exhibit stronger f0 perturbation in the speakers' productions.

Keywords: Mandarin Chinese, laryngeal contrast, aspiration, fundamental frequency (f0), production-perception relation, secondary cue

Edited by:
Georgia Zellou,
University of California, Davis,
United States

Reviewed by:
Jiayin Gao,
Ludwig Maximilian University of
Munich, Germany
James Kirby,
Ludwig Maximilian University of
Munich, Germany

***Correspondence:**
Harim Kwon
hkwon20@gmu.edu

[†]*These authors share first authorship*

Specialty section:
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Communication

Received: *14 March 2022*
Accepted: *13 May 2022*
Published: *17 June 2022*

Citation:
Guo Y and Kwon H (2022) Production and Perception of Mandarin Laryngeal Contrast: The Role of Post-plosive F0.
Front. Commun. 7:896013.
doi: 10.3389/fcomm.2022.896013

INTRODUCTION

F0 Perturbation

Laryngeal properties (such as voicing or aspiration) of onset plosives influence the fundamental frequency, or f0, at the onset of the following vowels. This phenomenon, commonly referred to as f0 perturbation, has been widely attested across languages, such as Cantonese (Francis et al., 2006; Luo, 2018), Dutch (Löfqvist et al., 1989), English

(House and Fairbanks, 1953; Lehiste and Peterson, 1961; Hombert et al., 1979; Ohde, 1984; Löfqvist et al., 1989; Hanson, 2009), French (Kirby and Ladd, 2016), German (Kohler, 1982; Hoole and Honda, 2011), Italian (Kirby and Ladd, 2016), Japanese (Gao and Arai, 2019), Khmer (Kirby, 2018), Mandarin (Xu and Xu, 2003; Luo, 2018), Russian (Mohr, 1971), Spanish (Dmitrieva et al., 2015), Thai (Gandour, 1974; Kirby, 2018), Vietnamese (Kirby, 2018), Xhosa (Jessen and Roux, 2002), Yoruba (Hombert et al., 1979), among others. The most commonly reported pattern shows that a (phonologically) voiced plosive has a lower post-plosive f0 than a (phonologically) voiceless one, although there are some notable patterns.

First, f0 perturbation occurs in so-called true voicing languages and in aspirating languages alike. That is, it seems less relevant whether the language contrasts prevoiced vs. voiceless unaspirated categories or unaspirated vs. aspirated categories. For example, both Spanish and English show similar f0 perturbation (Dmitrieva et al., 2015). This might be because English unaspirated plosives are phonologically voiced (Kingston and Diehl, 1994; Hanson, 2009). However, findings on languages with a three-way laryngeal contrast (prevoiced vs. unaspirated vs. aspirated) suggest that the difference between unaspirated and aspirated categories cannot entirely be reduced to phonological voicing. For example, Kirby (2018) examines Khmer, Vietnamese, and Thai, all with the three-way contrast, and finds that aspirated plosives are followed by a higher f0 than the unaspirated ones, at least for some speakers in all three languages. This provides evidence for the bona fide effects of consonantal aspiration (or the lack thereof) on the following f0.

Although the commonly reported pattern of f0 perturbation is voiceless (or aspirated) plosives having higher post-plosive f0 than voiced (or unaspirated) ones, this is not always the case. For example, Xu and Xu (2003) report that, in Mandarin, f0 is lower after aspirated plosives than after unaspirated plosives because aspiration causes the sub-glottal air pressure to decrease sharply, lowering f0 at the release of the plosives. However, Luo (2018) provides contradicting findings such that aspiration in Mandarin raises f0 quite robustly. The cause of this discrepancy is unclear (see more in the section: F0 Perturbation in Mandarin).

Second, f0 perturbation is attested in both tonal languages and non-tonal languages although the effects are less robust in tonal languages. For instance, the f0 differences between English unaspirated and aspirated series can last more than 100 ms after the voicing onset whereas they last 40∼60 ms in a tonal language, Yoruba (Hombert et al., 1979). Other studies on tonal languages (e.g., Chen, 2011, on Shanghainese; Gandour, 1974, on Thai; Francis et al., 2006, on Cantonese; Xu and Xu, 2003, on Mandarin) also suggest that f0 perturbation is limited to the very onset of the vowel and its exact duration is determined by the tonal contexts. Furthermore, Kirby (2018) reports that in Thai and Vietnamese, the perturbation effect is clearly observed in citation forms, but not in connected speech. This indicates that the effects of f0 perturbation may interact not only with tonal contexts but also with sentence-level prosody. See also Hanson (2009), Chen (2011), and Xu and Xu (2003), for similar effects in English, Shanghainese, and Mandarin, respectively.

Third, though the magnitude of the f0 perturbation is quite small (ranging 8–16 Hz in different languages, Table 1 in Coetzee et al., 2018), listeners use the f0 at the vowel onset to determine the preceding consonant's laryngeal category across different languages. English listeners, for instance, use f0 as a cue to consonant's laryngeal category not only when VOT, the phonetic property that is primarily responsible for the laryngeal contrast, is ambiguous (e.g., Whalen et al., 1990), but also when it is not ambiguous (e.g., Whalen et al., 1993). Even in tonal languages, in which f0 is primarily responsible to carry tonal information, and the perturbation, if any, is less consistent and temporally limited, post-plosive f0 influences listeners' perceptual judgments on onset plosive's laryngeal category. For example, Francis et al. (2006) report that falling f0 contours at the onset of a high-level tone signal aspirated plosives to Cantonese listeners and this perceptual pattern does not match the f0 patterns in Cantonese plosive productions. They claim that the use of post-plosive f0 as a consonantal cue, therefore, does not originate from the experience of hearing the covarying VOT and f0. Rather, Cantonese listeners' perception shows the influence of the language-independent, general auditory enhancing effects among different phonetic properties (Kingston and Diehl, 1994; Francis et al., 2006).

Despite the universality of the phenomenon, the source of f0 perturbation is controversial. Some have argued that f0 perturbation is a physiological or physical epiphenomenon of consonantal voicing or aspiration (e.g., Hombert et al., 1979; Löfqvist et al., 1989). Several different hypotheses have been offered on the exact mechanism of f0 perturbation. First, the aerodynamic hypothesis claims that voiced plosives differ from voiceless ones in how air pressure changes during and after their oral closure, leading to differing f0 after the release. In the case of voiced plosives, supraglottal air pressure gradually builds up during the closure because voicing requires a continuous airflow through the glottis. This results in a decrease in the trans-glottal air pressure difference, which in turn leads to a decrease in f0. On the other hand, voiceless plosives have a greater volume of airflow from subglottal to supraglottal cavities upon the release, resulting in faster vocal fold vibration (but see also Xu and Xu, 2003). Another hypothesis claims that f0 perturbation arises from the states of vocal folds during plosive voicing (e.g., Halle and Stevens, 1971; Löfqvist et al., 1989). During the plosive closure, the vocal folds remain slack for voiced plosives whereas they are stiff for voiceless plosives to halt the vibration. The tension of the vocal folds influences the f0 of the flanking vowels, such that slack vocal folds lower, and stiff vocal folds raise, the rate of their vibration. Still another hypothesis claims that f0 perturbation is due to the larynx height difference between the voiced and voiceless plosives (e.g., Honda, 2004). To allow for vocal fold vibration during the closure, the larynx is lower for voiced plosives than for voiceless ones. As the larynx height is usually positively correlated with f0, voiced plosives have lower post-plosive f0 than voiceless ones.

Despite the differences in their exact mechanisms, these hypotheses commonly suggest that the effects of plosive voicing (or voicelessness) on the following f0 are automatic and determined by the biomechanics of the larynx. In contrast, it has also been claimed that speakers actively induce the f0 differences to enhance the phonological contrast (e.g., Kingston and Diehl, 1994; Kingston, 2007). Under this phonological hypothesis, post-plosive f0 is not a mere by-product of sustaining voicing during the plosive closure or aspiration after the plosive release. Rather, speakers enhance the phonological laryngeal contrast by enhancing covarying phonetic properties. This results in the plosives of different laryngeal categories having distinct post-plosive f0, prolonged beyond the very beginning of the vowel. Therefore, this hypothesis can readily explain why the languages that contrast prevoiced and voiceless plosives (e.g., Spanish) and those contrasting aspirated and unaspirated plosives (e.g., English) show similar f0 perturbation patterns. In addition, in tonal languages, speakers would not enhance consonantal contrast using post-plosive f0 because f0 plays a central role in conveying lexical (or grammatical) information (Francis et al., 2006).

As pointed out in previous research (e.g., Chen, 2011; Hoole and Honda, 2011; Dmitrieva et al., 2015), these two views, automatic vs. phonological, are not incompatible with each other. In fact, it is possible that the biomechanical factors determine the connection between the voicing and f0, which serves as the resource for speakers to use as an enhancement strategy for plosive laryngeal contrast. Building on this previous conversation on f0 perturbation, this study asks how speakers of a tonal language use post-plosive f0 as a consonantal cue. Focusing on the relation between plosive aspiration and post-plosive f0, we investigate the production and perception of Mandarin word-initial plosives in different tonal contexts. The rest of the introduction will briefly review the relevant background on Mandarin and present the main questions for the two experiments.

F0 in Mandarin

Lexical Tones

Mandarin has four lexical tones, typically described as high-level (Tone 1), rising (Tone 2), low-dipping (Tone 3), and falling (Tone 4) (e.g., Xu, 1997; Duanmu, 2007). In this paper, tones are abbreviated as T1, T2, T3, and T4, and syllables produced with a specific tone are noted with a number added to the syllable. For example, /t^{h}a1/ refers to the syllable /t^{h}a/ with T1.

Xu (1997) describes the f0 contours of the four lexical tones as the following. T1 begins with a high f0 and maintains the same level through the entire vowel; T2 starts with a low f0, and then falls slightly until 20% into the vowel before rising throughout the rest of the vowel; T3, in citation form, begins with a low f0, falls to the lowest f0 at the midpoint of the vowel, and then rises sharply to the end of the syllable although the final rise is usually absent in non-prepausal positions; and T4 starts with a high f0, and then drops sharply from the 20% of the vowel until the end of the syllable. As f0 perturbation due to onset consonant is expected to be most distinct in the beginning of the vowel (adjacent to the onset consonant), two important aspects of these tones should be noted. First, T1 and T4 begin with a high f0 while T2 and T3 with a low f0. Second, T1 has the most static f0 contour and, in connected speech, T2 and T4 have more dynamic f0 contours than T3 during the first half of the vowel.

As for the physiological properties of Mandarin tones, studies have shown that larynx height is in general positively correlated with f0 (e.g., Hallé, 1994; Moisik et al., 2014). Specifically, the larynx is higher at the syllable onsets in T1 and T4 than in T2 and T3. However, Moisik et al. (2014) claim that the role of larynx height may be only facilitatory and, thus, the relation between larynx height and tones is not necessarily straightforward. This suggests that speakers may utilize different laryngeal settings (including larynx height, and vocal fold tension, among other things) to produce different tonal targets in Mandarin.

F0 Perturbation in Mandarin

Mandarin plosives are typically classified as voiceless unaspirated and voiceless aspirated, with aspiration as the primary distinction (Mandarin plosives are henceforth referred to as **unaspirated** and **aspirated** plosives). The language does not have voiced obstruents and, thus, the voiced consonants that can occur as a word onset are sonorants, such as /m n l w j/.

Inconsistent results have been reported on f0 perturbation in Mandarin (e.g., Xu and Xu, 2003; Luo, 2018; Chi et al., 2019). Xu and Xu (2003) suggest that aspiration is associated with low f0 although the specific pattern can be influenced by the tonal contexts. The lowering of aspiration is more robust in tones beginning with a low f0 (T2/T3, henceforth low-initial tones) than in those with a high f0 (T1/T4, high-initial tones). They attribute this pattern to the aerodynamics of the aspiration, which is characterized by a rapid outward flow of a large volume of air at the release of a plosive. This airflow, occurring between the release of oral closure and the glottal pulsing, lowers the subglottal air pressure for the aspirated plosives more than for the unaspirated ones, decreasing post-aspirated f0. The effects of aerodynamic force become even stronger when the intended pitch is low which is realized with slack vocal folds. Therefore, at the onset of low-initial tones, the vocal folds are slack and the f0 difference between aspirated and unaspirated series is enlarged.

By contrast, Luo (2018) reports that aspiration raises the f0, which extends longer in high-initial tones than in low-initial tones. In T2, they did not find a clear pattern of f0 perturbation. As for the source of this pattern (higher f0 after aspirated than unaspirated plosives), Luo mentions that aspiration is typically associated with high transglottal air pressure, elevated larynx, and stiff vocal folds, all of which raise the f0. On the other hand, she attributes the longer f0 perturbation in the high-initial tones than in low-initial tones to speakers' control (e.g., Kingston and Diehl, 1994). According to Luo (2018), in Mandarin, high-initial tones are more salient than low-initial ones both phonologically and perceptually. Phonologically, high-initial tones are more likely to be preserved in phonological processes, and perceptually, listeners are more accurate in perceiving high-initial tones. Assuming that tonal language speakers actively suppress the biomechanically-motivated automatic f0 perturbation to enhance the tonal contrast (e.g., Hombert et al., 1979; Francis et al., 2006), there is less need for this suppression when the tones are salient. Therefore, in Mandarin, high-initial tones allow for more f0 variability than low-initial tones.

The cause of the divergent findings in Xu and Xu (2003) and Luo (2018) is unclear. However, it is worth mentioning that the participants in both studies are all female speakers, who produce the target syllables embedded in different carrier phrases. In Luo's (2018) carrier phrase, the target syllables are always preceded by T1 whereas Xu and Xu (2003) use two different types of carrier phrases differing in the preceding syllable tones, T1 and T3. The two studies also differ in how they use f0 measurements in their analyses. While Xu and Xu's (2003) analyses are based on the raw f0 measured in Hz, Luo (2018) uses z-scored f0 normalized by speaker. The different patterns are possibly due to a great inter-speaker variation, as well.

Chi et al. (2019) compare two male speakers' glottal opening and oral airflow in aspirated and unaspirated plosives in T1. Their findings corroborate the possibility of the inter-speaker variation. One of the two tested speakers does not show the f0-aspiration covariation but shows faster oral airflow in aspirated than unaspirated plosives, especially when preceding a low vowel /a/ (Figure 5 in Chi et al., 2019). This speaker shows a negative relationship between the post-plosive f0 and oral airflow rate, suggesting that the post-plosive f0 decreases as the oral airflow rate increases, presumably for the consonant aspiration and a low vowel. This is consistent with the aerodynamic interpretation in Xu and Xu (2003). However, the other speaker does not show this airflow rate difference between aspirated and unaspirated plosives. And only this speaker tends to produce higher f0 for aspirated than unaspirated plosives, consistent with Luo's (2018) findings, although the f0 difference is not large enough to distinguish the aspiration contrast.

Despite the diverging patterns and potential individual variation, the previous findings commonly suggest that the f0 perturbation in Mandarin is fairly limited to the vowel onset. This is consistent with previous findings in other tonal languages (e.g., Hombert et al., 1979; Francis et al., 2006; Kirby, 2018).

Current Study

This study examines the role of post-plosive f0 as a secondary cue for Mandarin plosive laryngeal contrast in two experiments. We ask how the lexical tone mediates the f0 patterns in production, as well as the listeners' perceptual responses. The f0 at the vowel onset is expected to be influenced, interactively, by the lexical tone and the perturbation effects due to the onset consonants.

Experiment 1 examines the plosive production of Mandarin speakers to investigate the f0 patterns at vowel onset, influenced by the laryngeal category of the onset consonant, in CV syllables. The central questions for Experiment 1 are (1) how the aspiration (or the lack thereof) of the onset consonant changes the f0 at the onset of voicing following the onset consonant, and (2) how the tonal contexts influence the relation between consonant aspiration and f0 at voicing onset, if any. As mentioned above, the existing findings on the f0 perturbation in Mandarin are divergent and inconclusive (e.g., Xu and Xu, 2003; Luo, 2018; Chi et al., 2019). We aim to provide an additional set of empirical data, including both female and male speakers, on the f0 perturbation in Mandarin.

Experiment 2 examines Mandarin plosive perception. In Experiment 2, we specifically ask (1) whether the f0 differences between different laryngeal categories, if any, are used by Mandarin listeners as a cue to the onset aspiration, and (2) how the tonal contexts influence the listeners' use of f0 as a consonantal cue, if at all. It is still unknown whether Mandarin listeners use f0 as a secondary cue to the laryngeal contrast, to the best of our knowledge. Since f0 is the primary cue for lexical tones in the language, Mandarin listeners might not rely on the post-plosive f0 to determine the laryngeal category of the onset plosives. If Mandarin listeners do use the post-plosive f0 as a cue for the onset plosive, such an outcome may have different interpretations depending on the findings in Experiment 1. If the production patterns provide evidence for the perceptual patterns (i.e., if the listeners' behaviors reflect the robust patterns present in the speakers' production), the listeners' behaviors can be attributed to their native language experience. On the other hand, if the listeners associate post-plosive f0 with consonant aspiration in the absence of systematic f0 perturbation patterns in Mandarin productions, their perceptual behaviors could be attributed to the general auditory enhancing effects (Kingston and Diehl, 1994; Francis et al., 2006).

EXPERIMENT 1: PRODUCTION

Experiment 1 examines Mandarin speakers' plosive productions in CV syllables, asking how f0 at the vowel onset changes as a function of the laryngeal category of the onset consonant, in different tonal contexts.

Methods

Participants

Twenty-five native speakers of Mandarin Chinese (15 female and 10 male, mean age = 26, range = 19∼46) were recruited from the George Mason University community, in Virginia, USA. They were self-identified as native speakers of Mandarin, born and raised in the North China. All participants learned and spoke English as their second language, but they reported to be dominant in Mandarin. The participants moved to the US at the mean age of 22 (range 19∼35) and had lived in the US for 1∼48 months (mean = 13) at the time of testing, except for one participant (F05), who had been in the U.S. for 20 years. After confirming the data from this participant were not distinct from the rest of the group, we decided to include her in the analysis. The individual data are provided in the **Supplementary Materials**. No participants reported any history of speech or hearing disorders. The participants received monetary compensation for their participation.

Stimuli

The stimuli were 24 monosyllabic Mandarin words, with 3 onset consonants (aspirated, unaspirated, sonorant) * 2 vowel contexts (low [a], high [u]) * 4 lexical tones. We were mainly interested in comparing aspirated and unaspirated plosives, and sonorant onsets were also included as fillers. For the onset consonants, we used /t/, /t^h/, and /w/, as they yielded the least number of lexical gaps when combined with the vowels /a/ and /u/. However, /t^{h}a2/ is still lexically missing in Mandarin and, thus, was substituted with /p^{h}a2/, as f0 patterns for /t^{h}a2/ and /p^{h}a2/ are known to be similar (Ohde, 1984; Xu and Xu, 2003). In order to avoid directly comparing syllables with different onsets, we also substituted /ta2/ with /pa2/.

Written Mandarin words corresponding to each of the 24 syllables were selected based on the word frequency data from the Modern Chinese Balanced Corpus (Xiao, 2010, corpus size = 100 million words). Only the words labeled as "most common" were selected. None of the selected words was a bound morpheme in Mandarin. For the complete list of stimuli, see **Appendix A**.

The selected words were embedded in a carrier phrase 请说___一次 (/t͡ɕʰiŋ3 ʂwɔ1 ____ ji2 t͡sʰi4/, 'Please say ____ one time.')[1], and visually presented to the participants. The visual prompts included the entire carrier phrase in Chinese characters, with the stimulus word both in Chinese characters and Pinyin[2].

Procedure

The experiment took place in a sound-attenuated booth at the Phonetics and Phonology Lab at George Mason University. Participants were seated in front of a Macbook computer that presented the stimuli. Their productions were digitally recorded onto a separate Macbook Pro, using a lapel microphone (Røde smartLav+) and an external Focusrite Scarlette Solo 2nd Generation audio-interface, with a sampling rate of 44.1 kHz via the Praat program (Boersma and Weenink, 2020). The microphone was attached to the participants' shirt on the upper chest, ∼6 inches away from the speakers' mouth.

The visual prompts for stimuli were presented to the participants one at a time in the middle of the laptop screen using PsychoPy (Peirce, 2007). In order to elicit a comparatively stable speaking rate across participants, the sentences were presented with a fixed inter-stimulus interval of 3.5 seconds. Participants were instructed to read aloud each sentence on the laptop screen as naturally as possible. All written and oral instructions were provided in Mandarin.

Each stimulus (24 words) was repeated 6 times in randomized orders, resulting in a total of 144 trials per speaker. The 144 trials were presented in two blocks of 3 repetitions, with a self-paced break between the blocks. Beforehand, a short practice block with 2 trials was included to familiarize the participants with the task. The recording session took approximately 10 minutes.

Measurements and Data Preparation

All measurements were taken using Praat (Boersma and Weenink, 2020) by one of the authors (YG). Before taking the measurements, 23 of 3,600 (144 tokens * 25 speakers, 0.6%) tokens were removed due to production errors (e.g., not producing the target word, hesitation, self-correction, unintended noise such as coughing or clearing throat, etc.). For the remaining tokens, three different acoustic landmarks were labeled for each target token with the stop onset: (1) the onset of the stop burst, (2) the onset of the periodicity of the vowel following the stop consonant, and (3) the offset of the vowel second formant. VOT was calculated by subtracting (1) from (2), and the vowel duration by subtracting (2) from (3). For the fillers with the sonorant onset, the segmentation between the approximant onset /w/ and the following vowel was determined by visual inspection of the spectral patterns. The boundary was located at the point where the second formant (F2) moved up from the steady-state (Peterson and Lehiste, 1960), as well as the amplitude increased suddenly. The higher formants were used when F2 was not useful.

The f0 values from 20 equidistant points of the post-onset vowel, and then the first 8 (out of 20) f0 values (from the first 35% of the vowel) were used in the subsequent statistical analyses. As the duration of Mandarin sentence-medial vowels varies according to the lexical tones (e.g., Deng et al., 2006), the absolute duration of the 35% of the vowel used in this time-normalized method differs across the tones (mean duration for T1 75 ms; T2 80 ms; T3 75 ms; and T4 71 ms)[3]. The f0 values were extracted using a Praat script, with a 600 Hz pitch ceiling, a 75 Hz pitch floor, and a 10 ms time step. Any tracking errors were hand-corrected. In this process, an additional 5.3% of the data were removed due to unreliable f0 tracking when the vowel was not modal-voiced. A large portion of these excluded data was due to creaky voice, mostly in T3, but to a smaller extent in the other tones, when the f0 was low (see Kuang, 2017, for the discussion on creaky voice in different Mandarin tones).

Results

All statistical analyses in this study were conducted in R (R Core Team., 2021). To investigate the f0 perturbation in different tonal contexts, we built a series of linear mixed-effects models using the *lme4* package (Bates et al., 2014) on the normalized f0 (z-score). Z-scores were used instead of the raw f0 values (Hz), to facilitate comparisons across different speakers. In the initial model, we included the following factors as the fixed effects: ONSET (aspirated, unaspirated, sonorant), lexical TONE (T1, T2, T3, T4), VOWEL height (low, high), TIME points (eight categories from 0 to 7), and their interactions. TIME was coded using the orthogonal polynomial coding scheme and the rest of the fixed factors were Helmert-coded. The random effects structure of the model was determined using a forward best path algorithm (Barr et al., 2013), and the final model included by-SUBJECT random intercept, as well as by-SUBJECT slopes for ONSET, TONE, and VOWEL. The best fitting model was selected by comparing models using the likelihood ratio tests. The interactions ONSET * TONE * VOWEL * TIME and TONE * TIME * VOWEL did not improve the model fit [$\chi^2 = 10.60$, $p = 0.99$; $\chi^2 = 15.90$, $p = 0.78$, respectively] and, thus, they

[1] Note that the post-plosive f0 is likely affected by the preceding T1 in the carrier sentence (see, for example, Xu and Xu, 2003, for the discussion on this carryover effects). According to Xu and Xu (2003), both f0-ASP and f0-UNASP are higher after T1/T4 than after T2/T3, but f0-UNASP shows greater carryover effects than f0-ASP. If this is the case, it is possible that the preceding T1 elevated f0-UNASP more than f0-ASP and, consequently, the difference between f0-ASP and f0-UNASP in the current outcome is overplayed in low-initial tones but underplayed high-initial tones.

[2] Pinyin was included because this experiment was designed in parallel with a separate study testing L2 learners of Mandarin. Native Mandarin speakers would not need Pinyin to read common words in Chinese.

[3] We also tried a different method, in which we extracted the f0 values every 8 ms for the first 64 ms of the post-onset vowel, but the results were consistent with those obtained from the time-normalized method reported here.

TABLE 1 | F0 difference (z-score): aspirated–unaspirated (Tukey HSD *post-hoc* pairwise comparisons).

Time points		0 (0%)	1 (5%)	2 (10%)	3 (15%)	4 (20%)	5 (25%)	6 (30%)	7 (35%)
Tone	Vowel								
T1	Low	0.11^{***}	0.15^{***}	0.18^{***}	0.19^{***}	0.18^{***}	0.15^{***}	0.15^{***}	0.12^{***}
	High	0.27^{***}	0.23^{***}	0.20^{***}	0.17^{***}	0.16^{***}	0.15^{***}	0.13^{***}	0.13^{***}
T2	Low	-0.20^{***}	-0.13^{***}	$-0.06^{(*)}$	−0.02	0.00	0.01	0.02	0.03
	High	0.02	0.01	0.02	0.03	0.04	0.07^{*}	0.07^{*}	0.07^{*}
T3	Low	-0.32^{***}	-0.26^{***}	-0.17^{***}	-0.14^{***}	-0.11^{***}	-0.12^{***}	-0.09^{**}	-0.08^{*}
	High	-0.14^{***}	-0.16^{***}	-0.14^{***}	-0.13^{***}	-0.11^{***}	-0.10^{**}	-0.09^{**}	-0.07^{*}
T4	Low	0.08^{**}	0.10^{***}	0.09^{**}	$0.06^{(*)}$	0.03	−0.04	−0.06	-0.09^{**}
	High	0.29^{***}	0.23^{***}	0.15^{***}	0.09^{**}	0.05	0.00	−0.03	$-0.06^{(*)}$

Significance codes: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, and $^{(*)}$ for $p < 0.1$. Shaded cells indicate significant f0 differences that are unidirectional starting from time point 0 and continuing without a break.

were discarded. Consequently, the best model included four predictors ONSET, TONE, VOWEL, and TIME with the three-way interactions ONSET * TONE * TIME, ONSET * TIME * VOWEL, and ONSET * TONE * VOWEL. The outcome of this final model is in **Appendix B (Table B1)**.

Here, we present *p*-values for each significant factor and interaction obtained from the likelihood ratio tests comparing the best model and the model without the factor/interaction under consideration. Significant interactions were followed by *post-hoc* analyses using Tukey's HSD tests using the *emmeans* package (Lenth, 2020). If a predictor is significant in multiple interactions (or a main effect and interactions), only the highest-level interaction is reported along with the results of *post-hoc* testing.

We found the following significant interactions: ONSET: TONE: VOWEL [$\chi^2 = 2843.1$, $p < 0.0001$], ONSET: TONE: TIME [$\chi^2 = 1667.5$, $p < 0.0001$], ONSET: TIME: VOWEL [$\chi^2 = 250.8$, $p < 0.0001$]. As the predictor of our main interest, ONSET, was involved in multiple three-way interactions, we conducted the *post-hoc* Tukey pairwise comparisons on ONSET * TONE * VOWEL * TIME. The results of the pairwise comparisons are summarized in **Tables 1**, **3**, **4**, using the differences between the β coefficient values of different onset consonants. Shaded in **Tables 1**, **3**, **4** are the cells with significant f0 differences presumably attributable to onset consonants – that is, the cells with unidirectional f0 differences starting from time point 0 (closest to the onset consonant) and continuing without a break.

Figure 1 presents the mean f0 contours of the post-onset vowels. The contours are smoothed with loess and the shading displays a 95% confidence interval. To facilitate the visual interpretation of the figure, the z-normalized f0 is converted back to the Hz scale using the group mean (Brunelle et al., 2020), and the f0 contours of the entire duration of the post-onset vowels are plotted instead of the first 35% used in the statistical analysis. The vertical dotted line is added to indicate the 35% threshold included in the statistical analysis. The f0 contours are time-normalized, aligned from the voicing onset to the vowel offset (see Xu and Xu, 2021, for the comparison between different alignments). Individual speakers' production data are presented in the **Supplementary Materials**.

Aspirated and Unaspirated Stops

The f0 contours following an aspirated plosive (f0-ASP) and those following an unaspirated plosive (f0-UNASP) showed distinct patterns, but both the direction and the duration of the f0 differences varied according to the tonal contexts (**Table 1**). As for the direction of the f0 differences, f0-ASP was higher than f0-UNASP (indicated by positive numbers in **Table 1**) in T1 and T4, while the pattern showed the opposite direction in T2 and T3 (with the exception for /t^hu2/∼/tu2/ pair which showed no significant difference). The perturbation duration was also mediated by the tonal contexts. Specifically, the longest perturbation duration was observed in T1 and T3. In T1, the f0-ASP differed significantly from the f0-UNASP throughout the selected 35% of the vowel in T1 and T3 (corresponding to the mean duration of 75 ms in both tones), followed by T4 (10∼15% or 20∼30 ms). The perturbation due to aspiration (or lack thereof) was fairly limited in T2, either to the vowel onset (5% or 11 ms) in the /a/ context or not significant in the /u/ context.

As for the effects of VOWEL, the syllables with the high vowel /u/ had higher f0 than those with the low /a/, showing the expected vowel-intrinsic f0 patterns (e.g., Whalen and Levitt, 1995). This effect of vowel-intrinsic f0 was greater in high-initial tones than in low-initial tones (see **Figure 1**). In addition, the difference between f0-ASP and f0-UNASP in high-initial tones was greater in the /u/-contexts than in the /a/-contexts, but the same difference in low-initial tones was greater in the /a/-contexts than in the /u/-contexts.

In addition, aspirated plosives had longer VOT than unaspirated plosives, as expected (**Figure 2**). The influence of plosive ASPIRATION (aspirated, **unaspirated**), lexical TONE (**T1**, T2, T3, T4), and VOWEL height (**low**, high) on VOT (ms) was examined in a linear mixed effect model (Bates et al., 2014). The reference levels are bold-faced. The model included the interactions among the fixed factors, and by-SUBJECT random intercept. The model output is presented in **Appendix B (Table B2)**. The results revealed a significant three-way interaction ASPIRATION * TONE * VOWEL, and the follow-up Tukey's HSD tests (Lenth, 2020) confirmed that aspirated and unaspirated stops were significantly different [$\beta = -97.7$, $p <$

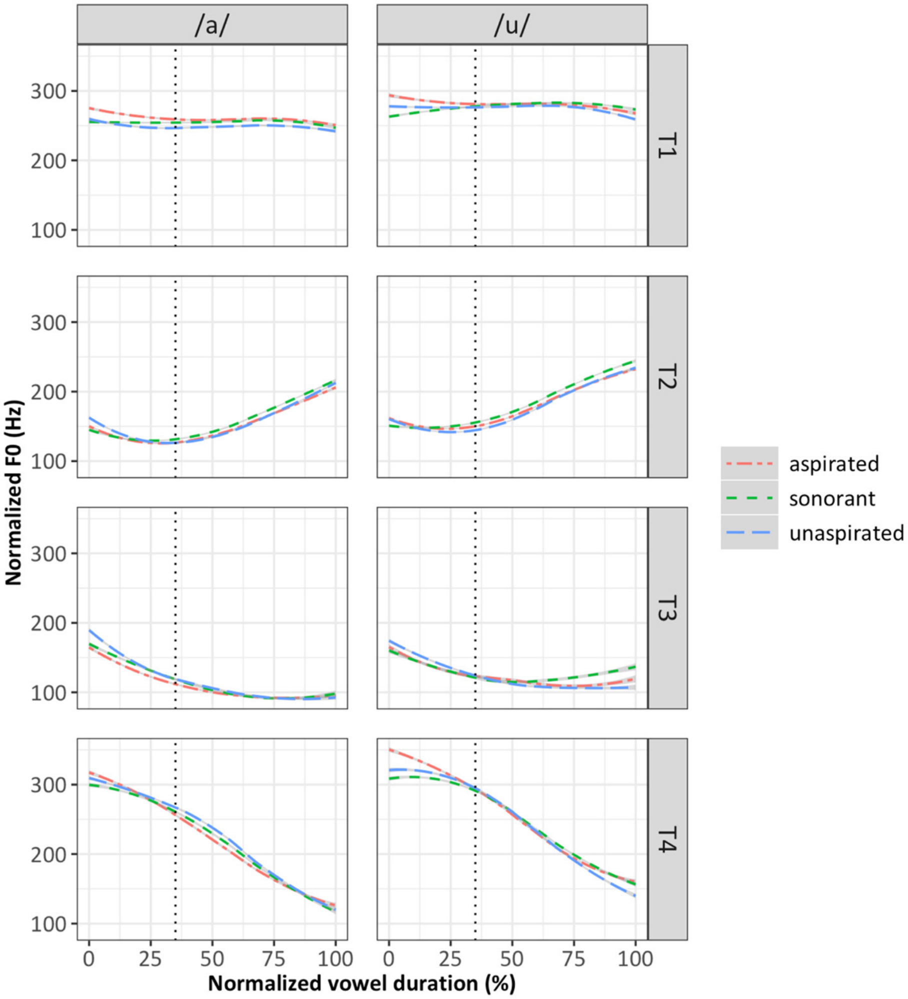

FIGURE 1 | Normalized F0 of Mandarin syllables.

0.0001]. Of interest to the current study, we also found significant effects of TONE on the VOT of aspirated plosives. As shown in **Figure 2**, the VOTs of aspirated plosives were the longest in T3, followed by T2, and T1 and T4 had the shortest VOT.[4] The results of the *post-hoc* Tukey's HSD comparisons are in **Table 2**. The VOTs of the unaspirated plosives did not show such effects of TONE.

Comparing Obstruents and Sonorants

Although the current study mainly aims to examine the f0 difference between aspirated and unaspirated plosives, we also compared f0-SON (f0 following a sonorant onset) with f0-ASP and f0-UNASP. Across different tone and vowel contexts, f0-UNASP was consistently greater than f0-SON, at least at the vowel onset (see **Table 3**). The difference between f0-ASP and f0-SON was less consistent (**Table 4**), varying mostly with the tonal contexts, in the same way as the difference between f0-ASP and f0-UNASP.

The duration of f0 perturbation varied in different tones as well as in different vowel contexts. The difference between f0-ASP and f0-SON mirrored the patterns showed between f0-ASP and f0-UNASP in high-initial tones. The difference between f0-ASP/UNASP and f0-SON also showed some influence of the vowel context. The difference lasted longer in the /u/-contexts than in the /a/-contexts in high-initial tones, but not in low-initial tones.

Interim Summary and Discussion

To summarize, the difference between f0-ASP and f0-UNASP showed opposite directions in high-initial tones and low-initial tones. On the other hand, the most consistent f0 difference across different tonal contexts was observed between f0-UNASP and f0-SON such that f0-UNASP was consistently higher than f0-SON. These outcomes suggest aspiration and voicing (or lack of voicing and aspiration) separately influenced the f0 at the vowel onset.

[4] Note our stimuli for T2 included bilabial /p^{h}a2/ and /pa2/ instead of /t^{h}a2/ and /ta2/. As coronal plosives usually have longer VOTs than labial plosives, this is expected to influence the reported VOT values for T2. We suspect that the VOT difference between T2 and T3 would have been exaggerated due to this difference in places of articulation.

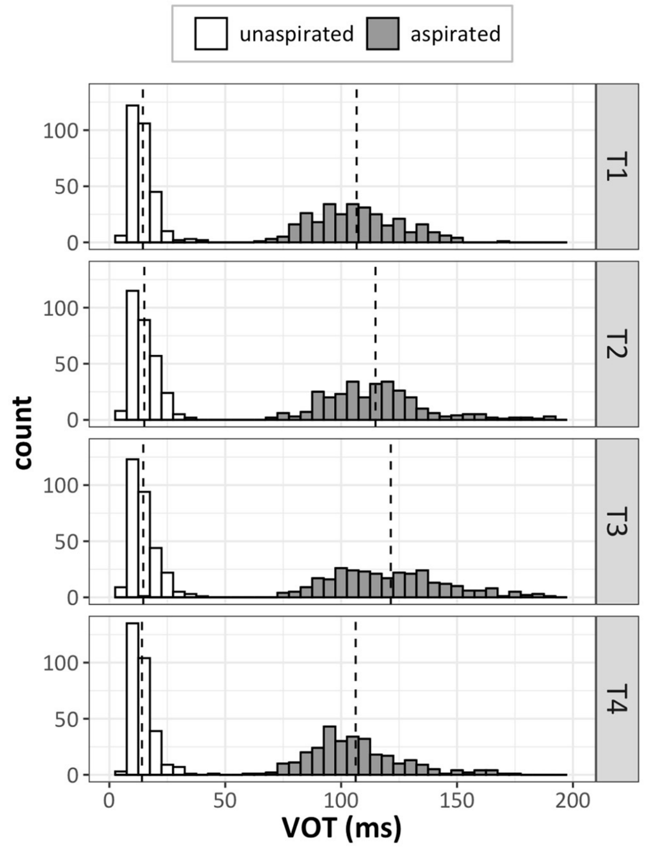

FIGURE 2 | Distribution of VOT across different tones. Dashed lines represent the mean VOT values.

TABLE 2 | Aspirated plosives' VOT in different tonal contexts (Tukey HSD *post-hoc* comparisons).

Tonal contrast	Estimate (β)	df	t ratio	*p*-value
T1–T2	−8.114	2347	−7.271	<0.0001***
T1–T3	−14.730	2347	−13.244	<0.0001***
T1–T4	0.494	2347	0.444	0.9708
T2–T3	−6.616	2347	−5.928	<0.0001***
T2–T4	8.608	2347	7.701	<0.0001***
T3–T4	15.224	2347	13.666	<0.0001***

Significance codes: *** for p < 0.001, ** for p < 0.01, * for p < 0.05, and $^{(*)}$ for p < 0.1.

First, aspirated plosives, compared to unaspirated plosives, influenced the f0 in different directions in high- vs. low-initial tones. Among the voiceless plosives, aspiration cooccurred with high f0 in the high-initial tones but with low f0 in the low-initial tones. The duration of this aspiration effect also depended on the tonal context. The difference between f0-ASP and f0-UNASP in the current study lasted the longest in T1 and T3, followed by T4. T2 showed little, if any, perturbation due to aspiration.

Second, although our main goal was to examine the perturbation due to consonant aspiration, we could also observe the voicing effect. F0-SON was consistently lower than f0-UNASP, suggesting that voicelessness raised (or voicing lowered) post-onset f0, consistent with the commonly observed cross-linguistic pattern. This effect was consistent throughout all tones.

The difference between f0-ASP and f0-SON seemed to reflect the interaction of these two effects. That is, if the f0-SON could be considered as the baseline, voicelessness (both unaspirated and unaspirated) raised f0, and in low-initial tones, aspiration lowered f0, resulting in little difference between f0-ASP and f0-SON. On the other hand, in high-initial tones, both aspiration and voicelessness raised f0, leading to a greater difference between f0-ASP and f0-SON.

The effect of vowel height interacted with the tonal contexts such that in high-initial tones, syllables with a high vowel showed greater f0 perturbation than those with a low vowel; in low-initial tones, syllables with a low vowel showed greater perturbation effects.

EXPERIMENT 2: PERCEPTION

Although complicated, the observed f0 perturbation patterns in Experiment 1 can be predicted as a function of the consonant's laryngeal category and the lexical tone. In this regard, the findings from Experiment 1 suggest that Mandarin has a systematic f0 perturbation at least for the tested speakers. Experiment 2 examines the perception of plosive aspiration contrast by the same Mandarin speakers. The purpose is to investigate whether Mandarin speakers, who produce systematically different f0 contours after aspirated and unaspirated plosives, use the f0 information to perceive the plosives' laryngeal categories.

Methods

Participants

The same individuals from Experiment 1 also participated in Experiment 2. Related to the task of Experiment 2, all participants reported to be right-handed.

Stimuli

Perception stimuli were created by recording natural productions of the syllables /t^{h}u/ in isolation, and manipulating them in Praat (Boersma and Weenink, 2020) to create a series of stops covarying in VOT and f0. A female native Mandarin speaker recorded the base syllables in four tones (i.e., /t^{h}u1/, /t^{h}u2/, /t^{h}u3/, /t^{h}u4/) in isolation. Aspirated stops were selected as the base tokens and unaspirated tokens were created by removing the aspirated portions from the base tokens. Consistent with previous studies using similar methods (e.g., Francis et al., 2006), removing the aspiration noise and shortening the VOT resulted in more natural sounding tokens than adding in aspiration noise and lengthening the VOT in our pilot works. The high back vowel /u/ was selected because /u/ provides a full set (all four tones) of real Mandarin words for both aspirated and unaspirated alveolar stops. We wanted to avoid the situation in which one of the choices is a word and the other is not. In addition, the vowel contexts did not influence the results in our pilot works using both /a/ and /u/ vowels.

TABLE 3 | F0 difference (z-score): unaspirated–sonorant (Tukey HSD *post-hoc* pairwise comparisons).

Time points		0 (0%)	1 (5%)	2 (10%)	3 (15%)	4 (20%)	5 (25%)	6 (30%)	7 (35%)
Tone	Vowel								
T1	Low	0.13***	0.05	−0.02	−0.05	−0.06$^{(*)}$	−0.06$^{(*)}$	−0.08*	−0.07*
	High	0.17***	0.11***	0.09**	0.05	0.02	0.00	−0.01	−0.02
T2	Low	0.24***	0.15***	0.06$^{(*)}$	0.02	−0.01	−0.04	−0.06$^{(*)}$	−0.07*
	High	0.14***	0.07*	0.02	−0.03	−0.07*	−0.12***	−0.13***	−0.16***
T3	Low	0.27***	0.20***	0.12***	0.09**	0.08*	0.07$^{(*)}$	0.05	0.04
	High	0.22***	0.17***	0.13***	0.10**	0.07*	0.05	0.03	0.00
T4	Low	0.16***	0.09**	0.03	0.02	0.02	0.03	0.03	0.03
	High	0.24***	0.19***	0.17***	0.15***	0.13***	0.13***	0.13***	0.12***

Significance codes: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, and $^{(*)}$ for $p < 0.1$. Shaded cells indicate significant f0 differences that are unidirectional starting from time point 0 and continuing without a break.

TABLE 4 | F0 difference (z-score): aspirated–sonorant (Tukey HSD *post-hoc* pairwise comparisons).

Time points		0 (0%)	1 (5%)	2 (10%)	3 (15%)	4 (20%)	5 (25%)	6 (30%)	7 (35%)
Tone	Vowel								
T1	Low	0.23***	0.21***	0.17***	0.14***	0.12***	0.08**	0.07*	0.06$^{(*)}$
	High	0.44***	0.35***	0.29***	0.22***	0.18***	0.15***	0.12***	0.10**
T2	Low	0.04	0.02	0.00	0.00	−0.02	−0.04	−0.04	−0.04
	High	0.16***	0.08*	0.04	0.00	−0.03	−0.05	−0.07$^{(*)}$	−0.08*
T3	Low	−0.05	−0.06	−0.05	−0.05	−0.03	−0.05	−0.04	−0.04
	High	0.08*	0.01	0.00	−0.03	−0.04	−0.05	−0.06	−0.07$^{(*)}$
T4	Low	0.25***	0.19***	0.12***	0.08*	0.04	−0.01	−0.03	−0.06
	High	0.53***	0.41***	0.31***	0.24***	0.18***	0.13***	0.11***	0.06$^{(*)}$

Significance codes: *** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, and $^{(*)}$ for $p < 0.1$. Shaded cells indicate significant f0 differences that are unidirectional starting from time point 0 and continuing without a break.

To obtain a fine-grained picture of the respective roles of VOT and f0 in the perception of Mandarin stop aspiration, 49 distinct syllables were initially created from each of the four base tokens (i.e., /t^{h}u1/, /t^{h}u2/, /t^{h}u3/, /t^{h}u4/). The 49 syllables covaried in stop VOT and post-stop f0, by fully crossing 7 steps of VOT and 7 steps of post-stop f0.

The mean VOT duration of the 4 base tokens was 99 ms, and the VOT step size was approximately 14 ms. Starting at the nearest zero crossing point from the end of the stop burst, about 14 ms of aspiration was manually removed incrementally in Praat until the VOT of the base token was around 14 ms. As a result, mean VOT values for each step were as follows: step 1 = 14 ms, step 2 = 28 ms, step 3 = 42 ms, step 4 = 56 ms, step 5 = 72 ms, step 6 = 86 ms, and step 7 = 99 ms.

Post-plosive f0 was manipulated using the TD-PSOLA (Moulines and Charpentier, 1990) implemented in Praat. First, the first 35% of the vowel was selected, and then the pitch curve of the selected vowel portion was simplified with the stylize function in Praat (frequency resolution 2 Hz). The onset f0 for each of the base tokens before manipulation were T1 = 323 Hz, T2 = 241 Hz, T3 = 210 Hz, and T4 = 371 Hz. Then, to create the 7 steps of post-plosive f0, the initial pitch point was either raised or lowered by 20 Hz, 40 Hz, and 60 Hz. F0 during the rest of the 35% of the vowel was proportionately increased or decreased. All the tokens were resynthesized with TD-PSOLA after the manipulation.

The tokens after manipulation were checked by four Mandarin native listeners for their naturalness, and all were judged to be good tokens of the original syllables. We conducted a pilot study with additional four Mandarin listeners, and VOT step 6 (84 ms) and step 7 (99 ms) never elicited different perceptual responses and, thus, VOT step 6 stimuli were removed from the experiment to keep the experiment short. The final set of perception stimuli included 168 (4 tones * 7 steps of f0 * 6 steps of VOT) unique tokens.

Procedure

Experiment 2, the perception experiment, was conducted after the production experiment, out of the concern that listening to the stimuli would influence the subsequent productions of the related sounds. After completing the production experiment, participants took a 5-min break before beginning the perception experiment.

Using PsychoPy (Peirce, 2007), the participants were presented with a forced-choice identification task. While listening to the stimuli, two Chinese characters constituting the aspirated and unaspirated pairs (e.g., 突/t^{h}u1/ vs. 督/tu1/) were

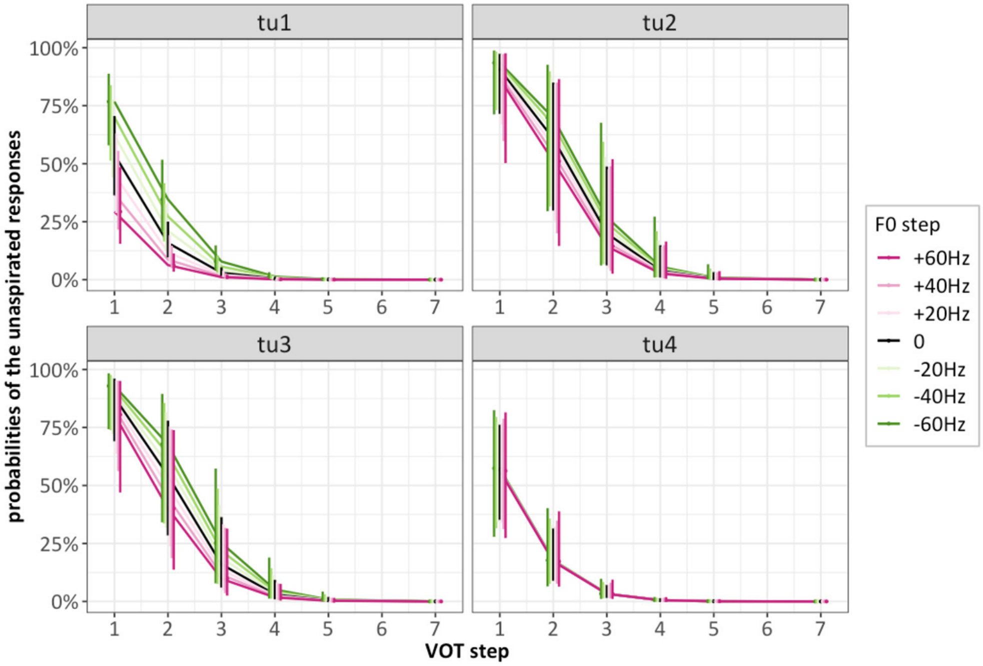

FIGURE 3 | Predicted perceptual responses based on the logistic regression model. Vertical lines represent the 95% confidence interval.

displayed on the laptop screen. Thirteen participants saw the screen with /t^h/- syllables on the left and /t/-syllables on the right, and 12 participants saw the opposite. The auditory stimuli were presented through Sennheiser HD 280 pro headphones. The participants were instructed to choose the word they heard by selecting one of the two characters using a Cedrus button box (model RB-740).

The experiment was blocked by the lexical tones and the order among the blocks was counter-balanced across participants. Within each block, each of the 42 tokens (7 f0 steps * 6 VOT steps) was repeated three times in different random orders. There were self-paced breaks between blocks. The entire task took about 20 minutes.

Results

A total of 12,600 responses (25 participants * 4 blocks * 42 tokens * 3 repetitions) were collected. Prior to the statistical analyses, the responses with the reaction time (measured from the onset of the audio stimuli to the button hit) that are more than 3 standard deviations away from the participant's mean (232 responses, 1.8%) were discarded. Then, to determine the influence of each acoustic property (VOT, post-plosive f0) on the identification of the onset laryngeal category, the responses (aspirated vs. unaspirated) were statistically analyzed using the binary logistic regression models built with the *lme4* packages in R (Bates et al., 2014). The reference category for the responses was aspirated and, thus, the coefficients β represent the log odds of unaspirated responses. The full model initially included VOT STEP, F0 STEP, TONE (T1, T2, T3, T4), and their interactions, as fixed effects. VOT STEP (1–7 without step 6) and F0 STEP (1-7) were included as continuous variables. TONE was orthogonally contrast coded (T1, T4 vs. T2, T3; T1 vs. T4; T2 vs. T3) to examine whether there are significant response differences between the high-initial tones (T1, T4) and the low-initial tones (T2, T3), as well as within the two tonal groups. The random effects structure of the model was determined using a forward best path algorithm (Barr et al., 2013), and the final model included by-SUBJECT and by-WORD intercepts, as well as by-SUBJECT slopes for VOT STEP, F0 STEP, and TONE. Interaction terms between fixed effects were included if they were directly related to our research question or if their inclusion improved the model fit based on a likelihood ratio test ($p < 0.05$). As a result, the final model included F0 STEP * TONE which was central to our research question. The full outcome of this final model is in **Appendix B (Table B3)**. A graph of predicted responses is in **Figure 3**. Raw response data for individual listeners are in the **Supplementary Materials**.

The likelihood ratio tests comparing the best model and the model without the predictor under consideration indicated that all fixed effects significantly influenced the listeners' responses. First, VOT STEP significantly contributed to model fit [$\chi^2 = 45.56, p < 0.0001$]. As shown in **Figure 3**, VOT step 1 elicited the highest rate of unaspirated responses across the four tones and, as the VOT increased, the possibility of unaspirated responses decreased. Second, the F0 STEP was also significant [$\chi^2 = 14.37$, $p = 0.0062$]: the higher the F0 STEP is, the less likely it is to elicit the unaspirated responses. Finally, as for TONE, the first tonal contrast (T1, T4 vs. T2, T3) contributed significantly to model fit [$\chi^2 = 30.47$, $p < 0.0001$], and high-initial tones (T1 and T4) elicited significantly less unaspirated responses than low-initial tones (T2 and T3) [$\beta = -3.82$, $p < 0.0001$]. The differences

TABLE 5 | Estimated trend of F0 step on tonal contrast (Tukey HSD *post-hoc* comparisons).

Tonal contrast	Estimate (β)	Standard Error	z. ratio	*p*-value
T1–T2	–0.206	0.134	–1.537	0.4152
T1–T3	–0.150	0.134	–1.122	0.6758
T1–T4	–0.323	0.141	–2.286	0.1013
T2–T3	0.055	0.125	0.442	0.9712
T2–T4	–0.118	0.134	–0.881	0.8148
T3–T4	–0.173	0.134	–1.293	0.5674

between the high-initial tones and the low-initial tones were the most conspicuous when VOT was short, as shown in **Figure 3**. For example, while the unaspirated responses were less than 50% at VOT step 2 in tones 1 and 4, in tones 2 and 3, a similar decrease was at step 3. This indicates that the stimuli belonging to the second step of VOT (28 ms), for instance, more likely elicited aspirated responses in the high-initial tones, but unaspirated responses in the low-initial tones. The second [$p = 0.74$] and third [$p = 0.35$] tonal contrasts were not significant, suggesting that listeners' responses in T1 vs. T4 and T2 vs. T3 were not significantly different.

The interaction F0 STEP: TONE was not significant [$\chi^2 = 5.37$, $p = 0.15$], but was included in the model as it was central to our research question. To verify whether the effects of F0 STEP across different tones, displayed in **Figure 3**, differed significantly, *post-hoc* Tukey tests were performed using the emtrends() function in the *emmeans* package (Lenth, 2020). It has been suggested that *post-hoc* analyses on non-significant interactions can be informative when the main effects of the predictors participating in an interaction are significant (e.g., Wei et al., 2012). The results of these *post-hoc* analyses suggest that the effects of F0 STEP did not differ as a function of TONE. None of the pairwise comparisons were significant, as shown in **Table 5**. Therefore, the current data do not provide evidence that the F0 STEP effects were influenced by tones. Rather, the current outcome appears to suggest that Mandarin listeners associated high post-plosive f0 with aspirated plosives across different tones.

Interim Summary and Discussion

The current findings demonstrate, as expected, that VOT is the primary cue of aspiration contrast in Mandarin. The unaspirated responses decreased as VOT became longer, across all f0 steps and lexical tones. At VOT step 1 (14 ms), which falls in the typical VOT range of the Mandarin unaspirated plosives (e.g., Rochet and Fei, 1991), the listeners provided the highest number of unaspirated responses, and starting from VOT step 4 (56 ms), the listeners tended to give mainly aspirated responses. The VOT categorical boundary for the aspirated-unaspirated plosives seemed to be different between high-initial tones vs. low-initial tones. Specifically, the VOT categorical boundaries occurred one step earlier in the high-initial tone stimuli than in the low-initial tone stimuli. At step 2, the low-initial tone stimuli yielded mostly unaspirated responses whereas the high-initial tone stimuli were more likely to yield aspirated responses (see **Figure 3**).

Although VOT was clearly the most influential cue for the aspiration, the listeners still used post-plosive f0 in deciding whether the plosive was aspirated or not. The current outcomes related to the f0 steps and lexical tones commonly suggest that the listeners associated high post-plosive f0 with the aspirated stops and low post-plosive f0 with unaspirated stops. The stimuli with raised f0 elicited more aspirated responses than those with lowered f0. In addition, stimuli with low-initial tones (T2, T3) elicited significantly more unaspirated responses than stimuli with high-initial tones (T1, T4). This is consistent with the pattern observed in the production experiment in which the aspirated plosives in T2 and T3 had longer VOT than those in T1 and T4 (**Figure 2**). This suggests that lower post-plosive f0, whether it be a part of the lexical tone or not, made the stops with an ambiguous VOT more likely to be judged as unaspirated than as aspirated.

Taken together, the current results suggest that Mandarin listeners extracted both consonantal and tonal information from f0 at the vowel onset. This perceptual pattern, however, did not precisely reflect the f0 perturbation observed in the same speakers' production patterns. In production, the difference between f0-ASP and f0-UNASP was not consistent across different tones, showing the opposite directions in high- vs. low-initial tones. Despite this divergent pattern in production, when VOT was ambiguous, the same speakers gave more aspirated responses in higher f0 steps both in high-initial and low-initial tones.

DISCUSSION

Post-onset F0 in Production

The main findings of Experiment 1, which compares f0-ASP, f0-UNASP, and f0-SON in four tonal contexts, can be summarized as the following. First, the difference between f0-ASP and f0-UNASP shows the opposite directions in high-initial tones and low-initial tones. In high-initial tones, f0-ASP is higher than f0-UNASP whereas f0-ASP is lower than f0-UNASP in low-initial tones. Second, f0-UNASP is consistently higher than f0-SON throughout the tonal contexts. Third, the difference between f0-ASP and f0-SON reflects the combination of these two effects. These outcomes suggest that the f0 at the vowel onset in Mandarin shows two separate perturbation effects, one due to aspiration and the other due to voicing. Between aspirated and unaspirated voiceless plosives, f0-ASP is higher in high-initial tones and lower in low-initial tones than f0-UNASP. Between voiceless plosives and voiced sonorants, voicelessness raises (or voicing lowers) f0 across the tonal contexts. Consequently, the difference between f0-ASP and f0-SON is greater in high-initial tones than in low-initial tones.

The current findings on f0 perturbation due to aspiration are partially consistent with the conflicting previous findings on Mandarin. Our findings in the low-initial tones are in line with Xu and Xu (2003), showing that aspiration lowers the post-plosive f0, compared to f0-UNASP, in low-initial tones. At the same time, we also find that in high-initial tones, aspiration raises the post-plosive f0, again compared to f0-UNASP, and this outcome is consistent with Luo's (2018) findings. The raising effects of the consonantal aspiration in Luo (2018) are greater in high-initial tones, the lowering effects in Xu and Xu (2003) are greater in low-initial tones, and our data show both of these patterns. These findings, taken together, reaffirm the dichotomy between the high-initial and low-initial tones.

The exact source of this dichotomy is puzzling, but we suggest that the tonal dichotomy is consistent with the interpretation that the f0 perturbation due to aspiration in Mandarin is biomechanically motivated. The observed tonal dichotomy can be explained by the differences in the laryngeal settings utilized in different tones. According to Moisik et al. (2014), the larynx height in general is positively correlated with f0 in Mandarin tone productions. As the laryngeal setting influences the vocal fold tension (e.g., Honda et al., 1999; Moisik et al., 2014), in high tones, the larynx is usually raised and the vocal folds are stretched and stiffened whereas the larynx is lowered and the vocal folds are slackened in low tones. When vocal folds are stiffened, they are resistant to vibration (i.e., require a greater volume of air flowing more rapidly than slack folds), but once they are set to vibrate, they vibrate at a high frequency. Also, stiffer vocal folds are often accompanied by a narrower glottal opening during the voiceless portion of a plosive (e.g., McCrea and Morris, 2005; Narayan and Bowden, 2013). On the other hand, slackened vocal folds are more prone to vibration and a wide glottal opening during a plosive.

The difference in the status of the vocal folds and the glottis has two notable consequences in the current study. The first consequence is the VOT difference in high-initial vs. low-initial tones. In the current study, aspirated plosives in high-initial tones have shorter VOT than those in low-initial tones (see **Figure 2**). According to McCrea and Morris (2005) and Narayan and Bowden (2013), stiff vocal folds and a narrow glottal opening result in shorter VOT of aspirated plosives, presumably accompanied by a faster airflow, in high f0 environments than in low f0 environments. The second consequence is the influence of aspiration on the post-plosive f0. Depending on the laryngeal settings for different tones, the influence of plosive aspiration on the post-plosive f0 can take different forms. According to the aerodynamic predictions, as claimed in Xu and Xu (2003), aspirated plosives, with a greater volume of air escaping through glottis between the oral release and the voicing onset, have a lower subglottal air pressure than unaspirated plosives at the voicing onset. This results in the f0-ASP being lower than f0-UNASP. This pattern (f0-ASP < f0-UNASP) appears when the vocal folds are slack and the glottal opening is wider, as in the low-initial tones in Mandarin. We claim that, in the high-initial tones, the aerodynamic effect is manifested in a different form because of the high tension of the vocal folds. As stiff vocal folds are more resistant to vibration and require a faster airflow to vibrate, the subglottal air pressure would not go down as much even in aspirated plosives. That is, the laryngeal setting and the resulting vocal fold tension in the high-initial tones require a higher trans-glottal pressure threshold than those in the low-initial tones, to initiate phonation at the onset of voicing after the plosive release. If the subglottal air pressure were to go down to the same extent regardless of the vocal fold tension, the trans-glottal pressure difference would not have been enough for the stiff folds to vibrate in the high-initial tones. Consequently, in the high-initial tones, the faster airflow in aspirated plosives (than in unaspirated plosives, see also Klatt et al., 1968), when combined with the high tissue tension and the narrow glottal opening, would increase the f0-ASP more than f0-UNASP. Chen (2011) proposes a similar dichotomy (tense vocal folds in a high-f0 context and slackened vocal folds in a low-f0 context, giving rise to distinct f0 perturbation patterns) for cross-linguistic variation. Our findings suggest that the tonal dichotomy can be observed even within a language.

Finally, although we suggest that the f0 perturbation in Mandarin is attributable to the biomechanics of the larynx, the current findings are also consistent, in several different aspects, with the claim that speakers of tonal languages would control the f0 perturbation to enhance (or not to impede) the tonal contrast (e.g., Hombert et al., 1979; Francis et al., 2006). First, the magnitude of the perturbation is greater in high-initial tones than in low-initial tones. Assuming that the high tones are salient in Mandarin (Luo, 2018) and the tones that are already salient do not need to be further enhanced, Mandarin speakers have more room for f0 variation in high-initial tones than in low-initial tones. This, according to Luo (2018), is the reason why the f0 raising due to aspiration is greater in high-initial tones in her study. Our findings differ from Luo's (2018) that we observe not only the f0 raising in high-initial tones but also the lowering in low-initial tones. Still the size of the difference between f0-ASP and f0-UNASP is greater in high-initial tones than in low-initial tones (**Table 1**), consistent with the claim that speakers would restrict the biomechanically-motivated f0 fluctuations when the tonal contrast is less salient and, thus, more vulnerable to misperception. Second, the perturbation lasts longer in the tones with a static f0 contour during the first half of the vowel than in those with a dynamic f0 contour. In Mandarin, the f0 contours for T1 and T3 are relatively steady during the first half of the vowel whereas those for T2 and T4 are more dynamic (see the section: Lexical tones). And the current findings indicate that the difference between f0-ASP and f0-UNASP lasts the longest in T1 and T3, followed by T4, and then T2 (**Table 1**). This seems to provide evidence for the speakers' control over f0 perturbation in a (subconscious) effort to preserve the tonal contrast. When the tones require dynamic f0 changes earlier in the vowel, speakers suppress the f0 variation automatically induced by the onset consonant. Tones with relatively steady f0 contours, on the other hand, would allow for more variability in f0 due to non-tonal factors, such as the aspiration of onset consonants.

Post-onset F0 in Perception

Our production data show that the f0 perturbation in Mandarin varies according to the lexical tones. As discussed in Post-onset f0 in production, this variation appears to be systematic, reflecting different laryngeal maneuvers for different tonal targets. Still, the same speakers, when they are presented with auditory stimuli varying in plosive VOT and post-plosive f0, are more likely to select the aspirated category when the post-plosive f0 is high and when VOT is ambiguous. The associations (high f0-aspirated and low f0-unaspirated) are valid even in the low-initial tones which show an opposite perturbation pattern in the production. In other words, there seems to be an intriguing mismatch between the production and the perception with regard to Mandarin speakers' use of f0 as a cue for consonant aspiration.

We propose several different factors contributing to this apparent mismatch. First, listeners are more attentive to the phonetic patterns present in salient contexts. Since Mandarin high-initial tones are more salient than low-initial tones both phonologically and perceptually, as suggested by Luo (2018), listeners may use the pattern presented in the salient tones that associates high f0 with aspirated plosives even when they perceive the low-initial tone stimuli. The production patterns in less salient low-initial tones are likely to be unattended. Second, the distribution of Mandarin lexical tones also suggests that the perturbation patterns in high-initial tones are more prevalent in the language. Liu and Ma (1986), based on their survey of two different corpora, the National Standard Corpus of Mandarin Words and the Chinese Vocabulary Corpus, show that T4 is the most frequent (32%) and T3 is the least frequent (17%) in Mandarin. T1 and T2 account for 24∼25% of Mandarin words. This means that the two high-initial tones (T1 and T4) compose more than half (56∼57%) of the Mandarin lexicon while the two low-initial tones, when combined, comprise about 40% of the lexicon. In addition, T3 is subject to tone sandhi (Duanmu, 2007), and when followed by another T3, becomes T2, which has the minimal, if any, perturbation due to aspiration (see **Table 1**, and also the same pattern is reported in Luo, 2018). Taking all these together, Mandarin listeners are presumably exposed to the f0 perturbation pattern that f0-ASP is higher than f0-UNASP more frequently than to the opposite pattern. Also, even in the infrequent cases when the listeners actually hear the pattern of f0-ASP < f0-UNASP, they are less likely to attend to this covariation occurring in less salient tonal contexts. Therefore, we claim that the Mandarin listeners' perception reflects the predominant pattern in their production. The perturbation pattern from the low-initial tones (f0-ASP < f0-UNASP) is not robustly represented, as T3 is the least frequent in the language and vulnerable to sandhi, and the perturbation in T2 is weak at best. Consequently, Mandarin listeners are likely to learn, from their native language experience, that high f0 is associated with aspirated plosives and low f0 with the unaspirated plosives, and use the high post-plosive f0 as a secondary cue to consonant aspiration.

Francis et al. (2006) also report a discrepancy between production and perception, in their investigation of the f0 perturbation in Cantonese. Cantonese listeners use post-plosive f0 as a cue for consonant aspiration but Cantonese speakers' production does not provide evidence for the association between high f0 and plosive aspiration. As the listeners' perceptual responses cannot be explained by their native language experience, Francis et al. (2006) claim that the listeners' perception is guided by a language-independent, general auditory enhancing effects among different phonetic properties (e.g., Kingston and Diehl, 1994), which could have been facilitated by the listeners' experience with English. Unlike Francis et al. (2006), we do see evidence for the association between high f0 and aspiration in Mandarin speakers' production. This suggests that the perceptual pattern observed in the current study may not be entirely due to the general auditory effects but, rather, due to the listeners' native language experience. However, we acknowledge that we cannot rule out the potential influence of the English experience. The participants in this study speak English as their second language, residing in Virginia, USA at the time of testing. We still expect the English influence, if any, to be minimal since bilingual listeners' categorization, which requires language-specific phonological judgments, shows the language mode effects (e.g., Antoniou et al., 2012). In this study, the experiments were carried out in Mandarin by a native Mandarin-speaking experimenter, and the perception task asked the listeners to select the Mandarin character matching the stimuli they heard.

Concluding Remarks: Mandarin Aspiration Contrast

The current outcomes confirm that VOT is the phonetic property primarily responsible for Mandarin aspiration contrast. In production, Mandarin aspirated plosives and unaspirated plosives are well-separated by the VOT alone (**Figure 2**), and Mandarin listeners primarily rely on VOT to distinguish the aspirated plosives from the unaspirated ones in perception (**Figure 3**). The VOT boundary, however, seems to vary according to the tonal contexts. The VOT of aspirated plosives is greater in low-initial tones than in high-initial tones in production (**Figure 2**). We suggest that this variation arises from the biomechanics of the larynx as, in high f0 ranges, VOT of aspirated stops decreases due to vocal fold tension (McCrea and Morris, 2005; Narayan and Bowden, 2013). In perception, Mandarin listeners are sensitive to this contextual VOT variation, providing more unaspirated responses in low-initial tones than in high-initial tones (**Figure 3**). Taken together, these findings suggest that the VOT boundary for Mandarin aspiration contrast is flexible and influenced by the tonal contexts. This is comparable to the well-documented covariation between VOT and place of articulation. In production, labial plosives have the shortest VOT, with the plosives of backer places of articulation having longer VOT (e.g., Peterson and Lehiste, 1960; Cho and Ladefoged, 1999). And listeners attend to this systematic variation. For example, the VOT boundary between voiced and voiceless categories is at a lower VOT range in labial plosives than in velar plosives (e.g., Miller, 1977; Benkí, 2001). When the variation in the speech signal is systematic, although it may not be uniform across contexts, the contextual variation does not impede but facilitates listeners' perception.

The current findings also provide evidence for a systematic variation in post-plosive f0 influenced both by the consonant aspiration and by the lexical tone. Depending on whether the tone begins at a high vs. low f0 range, the consonantal influence on f0 takes a different form. This can be attributed to the different laryngeal settings for different tonal targets. Despite the variation in production, Mandarin listeners use the post-plosive f0 as a secondary cue for plosive aspiration, associating high f0 with the aspirated category even in the low-initial tones which show an opposite perturbation pattern in the production. When the stimuli VOT is within a typical range of aspirated or unaspirated plosives, the listeners' responses are predominantly determined by the stop VOT. However, when the VOT is ambiguous (step 2 in high-initial tones and step 3 in low-initial tones, **Figure 3**), high post-plosive f0 stimuli, in general, yielded more aspirated responses despite a fairly large inter-listener variation (see the individual data in the **Supplementary Material**). That being said, the overall perceptual pattern pooled across the listeners may arguably originate from the f0 perturbation patterns in Mandarin production. As the high-initial tones are more salient and more prevalent in the Mandarin lexicon, the listeners attend more to the perturbation patterns present in high-initial tones (f0-ASP > f0-UNASP) than those in low-initial tones (f0-UNASP > f0-ASP). Although post-plosive f0 varies according to the tonal contexts in production, its role as the secondary cue to consonant aspiration in perception does not seem to be modulated by the tonal contexts.

Finally, the current study only reports the pooled results, but we should note that the data exhibit a considerable individual variation in both experiments (see the **Supplementary Material**). In production, some speakers show a quite clear f0 perturbation conforming to the group pattern while others show the conforming pattern only in a few tones but not in the others. In perception, post-plosive f0 does not seem to be an informative cue to consonant aspiration for all listeners, and some listeners seem to use f0 differently than others. The reason for these variations is unclear, and they do not seem to be structured in an immediately noticeable way. Still, this individual variation is intriguing and calls for a focused investigation, which we leave for a future study.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by George Mason University IRB. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YG: study conception and design, data collection and analysis, interpretation of results, and writing the initial draft. HK: supervising, data analysis, data visualization, interpretation of results, and writing and revising the manuscript. Both authors reviewed the results and approved the final version of the manuscript.

FUNDING

This work was supported by the George Mason University Dissertation Completion Grant awarded to YG and by the Linguistics Program at George Mason University.

ACKNOWLEDGMENTS

This work is a revised version of part of YG's dissertation (*Production and Perception of Laryngeal Contrasts in Mandarin and English by Mandarin Speakers*. George Mason University, 2020). The authors would like to thank Georgia Zellou and two reviewers for their constructive comments, the audience at the 178th meeting of the Acoustical Society of America and LabPhon 17 where portions of this work have been presented, Steven Weinberger, Vincent Chanethom, and Doug Wulf for helpful discussions related to this work, and participants for making this study possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.896013/full#supplementary-material

REFERENCES

Antoniou, M., Tyler, M. D., and Best, C. T. (2012). Two ways to listen: do L2-dominant bilinguals perceive stop voicing according to language mode? *J. Phone.* 40, 582–594. doi: 10.1016/j.wocn.2012.05.005

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1.1-7*. Available online at: http://CRAN.Rproject.org/package1/4lme4

Benkí, J. R. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *J. Phone.* 29, 1–22. doi: 10.1006/jpho.2000.0128

Boersma, P., and Weenink, D. (2020). *Praat: Doing Phonetics by Computer*, *Version 6.1.12*. Available online at: http://www.praat.org/

Brunelle, M., Tấn, T. T., Kirby, J., and Giang, Đ. L. (2020). Transphonologization of voicing in chru: studies in production and perception. *Lab. Phonol.* 11, 15. doi: 10.5334/labphon.278

Chen, Y. (2011). How does phonology guide phonetics in segment–f0 interaction? *J. Phone.* 39, 612–625. doi: 10.1016/j.wocn.2011.04.001

Chi, Y., Honda, K., and Wei, J. (2019). "Glottographic and aerodynamic analysis on consonant aspiration and onset f0 in Mandarin Chinese," in; *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brighton), 6480–6484.

Cho, T., and Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *J. Phone.* 27, 207–229. doi: 10.1006/jpho.1999.0094

Coetzee, A. W., Beddor, P. S., Shedden, K., Styler, W., and Wissing, D. (2018). Plosive voicing in afrikaans: differential cue weighting and tonogenesis. *J. Phone.* 66, 185–216. doi: 10.1016/j.wocn.2017.09.009

Deng, D. 邓丹, Feng, S. 石锋, and Lu, S. 吕士楠(2006). 普通话与台湾国语声调的对比分析[The contrast on tone between Putonghua and Taiwan Mandarin]. 声学学报*[Sheng Xue Xue Bao – Acta Acoustica]* 31, 536–541.

Dmitrieva, O., Llanos, F., Shultz, A. A., and Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in Spanish and English. *J. Phone.* 49, 77–95. doi: 10.1016/j.wocn.2014.12.005

Duanmu, S. (2007). *The Phonology of Standard Chinese*. New York, NY: Oxford University Press.

Francis, A. L., Ciocca, V., Wong, V. K. M., and Chan, J. K. L. (2006). Is fundamental frequency a cue to aspiration in initial stops? *J. Acoust. Soc. Am.* 120, 2884–2895. doi: 10.1121/1.2346131

Gandour, J. T. (1974). Consonant types and tone in Siamese. *J. Phone.* 2, 337–350. doi: 10.1016/S0095-4470(19)31303-8

Gao, J., and Arai, T. (2019). Plosive (de-)voicing and f0 perturbations in Tokyo Japanese: positional variation, cue enhancement, and contrast recovery. *J. Phone.* 77, 100932. doi: 10.1016/j.wocn.2019.100932

Halle, M., and Stevens, K. N. (1971). A Note on Laryngeal Features. Quarterly Progress Report, Research Laboratory of Electronics, MIT. 101, 198–213.

Hallé, P. (1994). Evidence for tone-specific activity of the sternohyoid muscle in modern standard Chinese. *Lang. Speech* 37, 103–123. doi: 10.1177/002383099403700201

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *J. Acoust. Soc. Am.* 125, 425–441. doi: 10.1121/1.3021306

Hombert, J. M., Ohala, J. J., and Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language* 55, 37–58. doi: 10.2307/412518

Honda, K. (2004). Physiological factors causing tonal characteristics of speech: From global to local prosody. *Proc Speech Prosody* 2004, 739–744.

Honda, K., Hirai, H., Masaki, S., and Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Lang. Speech* 42, 401–411. doi: 10.1177/00238309990420040301

Hoole, P., and Honda, K. (2011). "Automaticity vs. feature-enhancement in the control of segmental f0," in *Where do phonological features come from?: Cognitive, physical and developmental bases of distinctive speech categories Language Faculty and Beyond (LFAB): Internal and external variation in linguistics*, eds N. Clements and R. Ridouane (Amsterdam: John Benjamins), 133–171.

House, A. S., and Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am.* 25, 105–113. doi: 10.1121/1.1906982

Jessen, M., and Roux, J. C. (2002). Voice quality differences associated with stops and clicks in Xhosa. *J. Phone.* 30, 1–52. doi: 10.1006/jpho.2001.0150

Kingston, J. (2007). "Segmental influences on f0: automatic or controlled?" in *Tones and Tunes, Volume 2: Experimental Studies in Word and Sentence Prosody*, eds Gussenhoven and T. Riad (Berlin: Mouton de Gruyter), 171–201.

Kingston, J., and Diehl, R. L. (1994). Phonetic knowledge. *Language* 70, 419–494. doi: 10.1353/lan.1994.0023

Kirby, J. (2018). Onset pitch perturbations and the cross-linguistic implementation of voicing: Evidence from tonal and non-tonal languages. *J. Phone.* 71, 326–354. doi: 10.1016/j.wocn.2018.09.009

Kirby, J., and Ladd, D., R. (2016). Effects of obstruent voicing on vowel F0: evidence from "true voicing" languages. *J. Acoust. Soc. Am.* 40, 2400–2411. doi: 10.1121/1.4962445

Klatt, D. H., Stevens, K. N., and Meade, J. (1968). "Studies of articulatory activity and airflow during speech in sound production in man," in *Annals of the New York Academy of Science*, eds A. Bouhuys (New York, NY), 42–55.

Kohler, K. J. (1982). F0 in the production of lenis and fortis plosives. *Phonetica* 39, 199–218. doi: 10.1159/000261663

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *J. Acoust. Soc. Am.* 142, 1693–1706. doi: 10.1121/1.5003649

Lehiste, I., and Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *J. Acoust. Soc. Am.* 33, 419–425. doi: 10.1121/1.1908681

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means. R Package Version 1.4.5.* Available online at: https://CRAN.R-project.org/package=emmeans

Liu, L. Y. 刘连元, and Ma, Y. F. 马亦凡(1986). 普通话声调分布和声调结构频度[The distribution of Mandarin tones and the frequency of tonal phrases]. 语文建设 *[Language Planning]* 3, 21–23.

Löfqvist, A., Baer, T., McGarr, N. S., and Story, R. S. (1989). The cricothyroid muscle in voicing control. *J. Acoust. Soc. Am.* 85, 1314–1321. doi: 10.1121/1.397462

Luo, Q. (2018). *Consonantal Effects on F0 in Tonal Languages (Doctoral dissertation)*. Michigan State University, East Lansing.

McCrea, C. R., and Morris, R. J. (2005). The effects of fundamental frequency levels on voice onset time in normal adult male speakers. *J. Speech Lang. Hear. Res.* 48, 1013–1024. doi: 10.1044/1092-4388(2005/069)

Miller, J. L. (1977). Nonindependence of feature processing in initial consonants. *J. Speech Lang. Hear. Res.* 20, 519–528. doi: 10.1044/jshr.2003.519

Mohr, B. (1971). Intrinsic variations in the speech signal. *Phonetica* 23, 65–93. doi: 10.1159/000259332

Moisik, S. R., Lin, H., and Esling, J. H. (2014). A study of laryngeal gestures in Mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (SLLUS). *J. Int. Phon. Assoc.* 44, 21–58. doi: 10.1017/S0025100313000327

Moulines, E., and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* 9, 453–467. doi: 10.1016/0167-6393(90)90021-Z

Narayan, C., and Bowden, M. (2013). Pitch affects voice onset time (VOT): a cross-linguistic study. *Proc. Meet. Acoust.* 19, 060095. doi: 10.1121/1.4800681

Ohde, R. N. (1984). Fundamental frequency as an acoustic correlate of stop consonant voicing. *J. Acoust. Soc. Am.* 75, 224–230. doi: 10.1121/1.390399

Peirce, J. W. (2007). PsychoPy—psychophysics software in python. *J. Neurosci. Methods* 162, 8–13. doi: 10.1016/j.jneumeth.2006.11.017

Peterson, G. E., and Lehiste, I. (1960). Duration of syllable nuclei in English. *J. Acoust. Soc. Am.* 32, 693–703. doi: 10.1121/1.1908183

R Core Team. (2021). *A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Available online at: https://www.R-project.org

Rochet, B. L., and Fei, Y. (1991). Effect of consonant and vowel context on Mandarin Chinese VOT: production and perception. *Can. Acoust.* 19, 105.

Wei, J., Carroll, R. J., Harden, K. K., and Wu, G. (2012). Comparisons of treatment means when factors do not interact in two-factorial studies. *Amino Acids* 42, 2031–2035. doi: 10.1007/s00726-011-0924-0

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1990). Gradient effects of fundamental frequency on stop consonant voicing judgments. *Phonetica* 47, 36–49. doi: 10.1159/000261851

Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *J. Acoust. Soc. Am.* 93, 2152–2159. doi: 10.1121/1.406678

Whalen, D. H., and Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *J. Phone.* 23, 349–366. doi: 10.1016/S0095-4470(95)80165-0

Xiao, H. 肖航(2010). 现代汉语通用平衡语料库建设与应用[The construction and application of the general modern Chinese balanced corpus]. 华文世界 *[Chinese World]*. 106, 24–29.

Xu, C. X., and Xu, Y. (2003). Effects of consonant aspiration on Mandarin tones. *J. Int. Phon. Assoc.* 33, 165–181. doi: 10.1017/S0025100303001270

Xu, Y. (1997). Contextual tonal variations in Mandarin. *J. Phone.* 25, 61–83. doi: 10.1006/jpho.1996.0034

Xu, Y., and Xu, A. (2021). Consonantal f0 perturbation in American English involves multiple mechanisms. *J. Int. Phon. Assoc.* 149, 2877–2895. doi: 10.1121/10.0004239

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.