frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 26 September 2022
DOI 10.3389/fcomm.2022.864127

OPEN ACCESS

EDITED BY
Joseph V. Casillas,
Rutgers, The State University of New Jersey, United States

REVIEWED BY
Jessie S. Nixon,
University of Tübingen, Germany
Allard Jongman,
University of Kansas, United States

*CORRESPONDENCE
Roger Yu-Hsiang Lo
roger.lo@ubc.ca

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

RECEIVED 28 January 2022
ACCEPTED 23 August 2022
PUBLISHED 26 September 2022

CITATION
Lo RY-H (2022) The dual role of post-stop fundamental frequency in the production and perception of stops in Mandarin-English bilinguals. *Front. Commun.* 7:864127. doi: 10.3389/fcomm.2022.864127

The dual role of post-stop fundamental frequency in the production and perception of stops in Mandarin-English bilinguals

Roger Yu-Hsiang Lo*

Department of Linguistics, University of British Columbia, Vancouver, BC, Canada

In non-tonal languages with a two-way laryngeal contrast, post-stop fundamental frequency (F0) tends to vary as a function of phonological voicing in stops, and listeners use it as a cue for stop voicing. In tonal languages, F0 is the most important acoustic correlate for tone, and listeners likewise rely heavily on F0 to differentiate tones. Given this ambiguity of F0 in its ability to signal phonological voicing and tone, how do speakers of a tonal language weight it in production and perception? Relatedly, do bilingual speakers of tonal and non-tonal languages use the same weights across different language contexts? To address these questions, the cross-linguistic performances from L1 (first language) Mandarin-L2 (second language) English bilinguals dominant in Mandarin in online production and perception experiments are compared. In the production experiment, the participant read aloud Mandarin and English monosyllabic words, the onsets of which typified their two-way laryngeal contrast. For the perception experiment, which utilized a forced-choice identification paradigm, both the English and Mandarin versions shared the same target audio stimuli, comprising monosyllables whose F0 contours were modeled after Mandarin Tone 1 and Tone 4, and whose onset was always a bilabial stop. The voice onset time of the bilabial stop and the onset F0 of the nucleus were manipulated orthogonally. The production results suggest that post-stop F0 following aspirated/voiceless stops was higher than that following unaspirated/voiced stops in both Mandarin and English production. However, the F0 difference in English was larger as compared to Mandarin, indicating that participants assigned more production weight to post-stop F0 in English than in Mandarin. On the perception side, participants used post-stop F0 as a cue in perceiving stops in both English and Mandarin, with higher post-stop F0 leading to more aspirated/voiceless responses, but they allocated more weight to post-stop F0 when interpreting audio stimuli as English words than as Mandarin words. Overall, these results argue for a dual function of F0 in cueing phonological voicing in stops and lexical tone across production and perception in Mandarin. Furthermore, they suggest that bilinguals are able to dynamically adjust even a secondary cue according to different language contexts.

KEYWORDS

fundamental frequency, stop, Mandarin, English, bilingual, production, perception, cue weighting

1. Introduction

Speech sounds contrast on a multitude of continuous acoustic dimensions, with some dimensions being used as primary cues to a phonological contrast while others play a more secondary part. Following Toscano and McMurray (2010), I use the term *cue* to refer to any source of information that allows the perceiver to distinguish between different responses (e.g., the response might be whether the sound is an [i] or an [a]). An example that is often given in this connection is Lisker's (1986) finding that potential cues to word-medial voicing in English (e.g., *rapid* vs. *rabid*) include duration of the preceding vowel, duration of the closure, voice onset time (VOT), presence of vocal fold vibration during closure, burst amplitude, fundamental frequency (F0) going into and out of the closure, among others. However, the reverse—that an acoustic dimension can serve as a cue for multiple phonological contrasts—is also true but often less studied. For instance, formant frequency is not only an important cue for vowel quality, but the transition for a formant frequency band also cues the place of articulation for stop consonants (e.g., Liberman et al., 1954). Given this many-to-many mapping between phonological contrasts and acoustic dimensions, ambiguity about how speakers encode various cues for a contrast and how listeners infer potential contrasts from a cue naturally arises.

The current study explores this ambiguity from the perspective of both speech production and perception. Specifically, I am interested in (i) whether and how F0 is used by speakers of a tonal language to signal and perceive phonological *voicing* in stops, aside from lexical tone, and (ii) whether the use of F0 might be mediated by different language contexts. These two questions are addressed in tandem by comparing L1 (first language) Mandarin-L2 (second language) English bilinguals' performances in production and perception of Mandarin and English stops. The production task involves the participants reading aloud words with a stop in the onset position, while the perception part asks the participants to respond in a forced-choice identification task based on synthetic continua of both VOT and F0 values.

2. Background

2.1. Fundamental frequency as a cue to lexical tone

Similar to segments, lexical tones contrast on multiple acoustic dimensions, such as duration and intensity; however, F0 has long been established as the most important acoustic correlate for tonal distinctions, as far as Mandarin is concerned (Ohala, 1978). Indeed, the tone letters in the International Phonetic Alphabet are in their essence a discretized representation over a speaker's full pitch range, and the descriptions for lexical tones in Mandarin closely follow the F0 as they unfold over a syllable—Tone 1: high-level ˥, Tone 2: mid-rising ˧˥, Tone 3: low-dipping ˨˩˦, and Tone 4: high-falling ˥˩. Even though F0 is not the only dimension that covaries with each tone in production (Ho, 1976), and it is not the only dimension that listeners take advantage of when distinguishing tones (e.g., Blicher et al., 1990), it is the primary source that Mandarin users rely on to signal and extract information regarding tonal contrast (Gandour, 1978).

In this study, I restrict the scope to only Tone 1 and Tone 4 for both theoretical and practical considerations. On the theoretical side, Tone 1 and Tone 4 are the only two tones in Mandarin that start with the same phonological tonal register (i.e., both start with a high target), so listeners need to track the F0 trajectory, at least for the initial portion of a tonal contour initiated with a high register, to reliably tell these two tones apart. This is an important consideration for the design of the perception experiment, as will be explained in Section 2.2.2. Also, given that both Tone 1 and Tone 4 begin in the upper part of the pitch range, post-stop F0 behaviors, which will be discussed in the next section, should be more comparable across these two tones, as there is evidence suggesting that post-stop F0 is contingent on pitch height.

2.2. Fundamental frequency as a cue to stop voicing

2.2.1. Post-stop F0 in English

It has been observed that F0 in the vowel following a stop consonant tends to correlate with voicing distinctions cross-linguistically [e.g., Cantonese (Francis et al., 2006; Luo, 2018; Ren and Mok, 2021), English (House and Fairbanks, 1953; Lehiste and Peterson, 1961; Lea, 1973; Hombert, 1978; Hombert et al., 1979; Ohde, 1984; Hanson, 2009), French (Kirby and Ladd, 2016), German (Kohler, 1982), Japanese (Gao and Arai, 2018), Korean (Han and Weitzman, 1970; Jun, 1996), Mandarin (Howie, 1976; Xu and Xu, 2003; Chen, 2011; Luo, 2018; Guo, 2020), Russian (Mohr, 1971), Spanish (Dmitrieva et al., 2015), Thai (Gandour, 1974; Ewan, 1976), Xhosa (Jessen and Roux, 2002), Yoruba (Hombert, 1978)]. This phenomenon is commonly labeled as post-stop F0 perturbation, pitch skip, obstruent intrinsic F0, co-intrinsic pitch, or onset F0 perturbation. For English, whose six stops come in phonologically voiced-voiceless pairs: /b/-/p/, /d/-/t/, and /g/-/k/, it is well-established that F0 at vowel onset is significantly higher following phonologically voiceless stops than following phonologically voiced ones, regardless of the presence of actual vocal fold vibration (e.g., Abramson and Lisker, 1985; Dmitrieva et al., 2015). This type of patterning has led Kingston and Diehl (1994) to argue that post-stop F0 is not purely a result of intrinsic physiological dependencies between the articulatory

and/or aerodynamic properties and the production of degrees of prevoicing or voicing delay—instead, it is at least partially the result of controlled processes referring to the phonological status of the consonant series.

The perceptual consequences of post-stop F0 to the voicing contrast are also firmly established for English: a higher post-stop F0 tends to lead to more voiceless responses than a lower F0, especially when VOT is ambiguous (Whalen et al., 1990, 1993; Francis et al., 2006). Some authors have attributed the perceptual effects of post-stop on voicing decisions to the observation that a low F0 enhances the perceptual "voicedness" of a stop by highlighting the percept of low-frequency periodic energy in the proximity of the stop release (Kingston and Diehl, 1994; Kingston et al., 2008).

2.2.2. Post-stop F0 in Mandarin

With regard to the post-stop F0 perturbation effect in Mandarin, which has six stops coming in unaspirated-aspirated pairs: /p/-/p^h/, /t/-/t^h/, and /k/-/k^h/, the existing literature depicts a mixed picture, with conflicting results across studies. Both English and Mandarin have two phonological voicing classes, with the voiced / unaspirated class typically having a short-lag VOT (under 30 ms) and the voiceless / aspirated class having a long-lag VOT (above 30 ms). Based on this similar phonetic implementation, one would expect Mandarin to pattern with English in terms of post-stop F0 effects, that is, aspirated stops should have a higher post-stop F0 than unaspirated stops. Indeed, this is the pattern found by Chen (2011) and Luo (2018). Based on read speech from 15 female native speakers of Mainland Mandarin reading monosyllabic CV words containing all six stops inserted in a carrier phrase, Luo (2018) found that aspirated stops were associated with greater F0 perturbation (i.e., a higher F0) than unaspirated stops, with a mean F0 difference in the range of 11.67 Hz and 18.35 Hz, depending on the lexical tone. With a similar experiment design to that in Luo (2018), but with gender-balanced speakers (10 females and 10 males), Chen (2011) also reached the conclusion that vowels following an aspirated stop had a higher F0 than those following an unaspirated stop in Taiwan Mandarin (for females, the difference in F0 ranged from 2 Hz and 14 Hz; for males, the range was between 2.8 Hz and 8 Hz). This general pattern was also reported in a blog post by Liberman (2014), based on the data from the Mandarin Chinese Phonetic Segmentation and Tone corpus (Yuan et al., 2014). However, as Liberman (2014) did not conduct statistical tests on this set of data, it is not yet clear if the difference was statistically significant (across genders, the mean F0 difference was between 1.5 Hz and 5.7 Hz for the /p/-/p^h/ contrast, between 1.0 Hz and 3.5 Hz for the /t/-/t^h/ contrast, and between 2.8 Hz and 7.2 Hz for the /k/-/k^h/ contrast). Rather puzzlingly, a pattern that is opposite to the above generalizations was also observed in the work by Xu and Xu (2003), where they reported that it was *unaspirated* stops that triggered a higher F0 on the onset of the following vowel (with a mean F0 difference ranging between 5 Hz and 50 Hz), using production data from seven female native speakers of Mainland Mandarin pronouncing disyllabic words containing /ta/ and /t^{h}a/ embedded in a carrier phrase. Even more interestingly, a recent work from Guo (2020), which used as stimuli tonal syllables with onsets /t/, /t^h/, or /w/ and rimes /a/ or /u/ in the four lexical tones, showed that the direction of post-stop F0 perturbation depended on the tone, such that F0 was higher following an aspirated stop only in Tone 1 and Tone 4 (i.e., tones beginning with a high register) while the opposite pattern was observed for Tone 2 and Tone 3, both of which have a low initial register.

More broadly, the issue of post-stop F0 perturbation in Mandarin is related to the debate of whether there is a trade-off between post-stop F0 and tone, and of whether the existence of tone attenuates the degree of post-stop F0 difference. While there are some studies that provide a positive answer [e.g., Gandour (1974) for Thai and Hombert (1978) for Yoruba], larger magnitudes have also been reported in tonal languages [e.g., Phuong (1981) for Northen Vietnamese, Shimizu (1994) for Thai, Xu and Xu (2003) for Mandarin, and Francis et al. (2006) for Cantonese]. In the current study, the parallel production experiments in Mandarin as well as English allow us to address this debate from a bilingual perspective. That is, the production data in Mandarin and English enables a comparison of the degree of post-stop F0 difference across a tonal and a non-tonal language within the same speaker.

The perceptual contribution of post-stop F0 to the voicing contrast in Mandarin is substantially less studied. To my knowledge, Guo (2020) is the first to systematically study whether post-stop F0 is used by Mandarin speakers as a cue when tasked to distinguish the stop voicing contrast in Mandarin. Using a two-alternative forced choice (2AFC) paradigm, Guo (2020) showed that Mandarin speakers capitalized on post-stop F0 to decode consonantal voicing information. However, the identification experiment in her study only required the listener to distinguish aspirated vs. unaspirated stops in the context of the same lexical tone (i.e., the two alternatives in the 2AFC paradigm only differed in stop voicing but shared the same lexical tone), and so it is still unclear whether Mandarin listeners continue to use post-stop F0 as a cue for voicing when they have to extract tonal information from pitch at the same time. The design of the current perception experiment addresses this problem, as explained in Section 4.3.

2.2.3. Post-stop F0 and F0 contour

Given that post-stop F0 is embedded in the global F0 trajectory that also encodes tonal and intonational information, this section briefly reviews the interaction between post-stop F0 and F0 contour in English and Mandarin. In English production, Hanson (2009) examined the effects of obstruents on F0

contour in either a high, low, or neutral pitch environment by having participants read CVm syllables in carrier sentences. She found that, in a high-pitch environment, the initial F0 contour following a voiceless stop was raised relative to the baseline /m/, but following a voiced stop, it closely approximated the baseline. In a low-pitch environment, however, both voiceless and voiced stops raised the initial F0 contour. In Mandarin production, regardless of whether aspirated stops were found to lead to a higher post-stop F0 than unaspirated ones (e.g., Chen, 2011; Luo, 2018) or otherwise (Xu and Xu, 2003), visual inspection of the F0 trajectories in these studies suggests that both aspirated and unaspirated stops raised the initial F0 contour in all lexical tonal contexts.

With respect to perception, much less is known about how F0 contour affects the perceived phonological voicing of the initial stop. It is well established that listeners of both tonal and non-tonal languages are sensitive to changes in F0 in signaling sentential intonation or lexical tone (e.g., Gandour, 1983; Ma et al., 2006; Barnes et al., 2010; Liu and Rodriguez, 2012; Xu and Mok, 2012; Dilley and Heffner, 2013; Leung and Wang, 2020). For instance, Gandour (1983) asked listeners of tonal languages (Cantonese, Mandarin, Taiwanese, Thai) and a non-tonal language (English) to make direct paired-comparison judgments of tone dissimilarity. His results revealed that the direction dimension was more important than the height dimension for listeners of a tonal language vs. a non-tonal language. Leung and Wang (2020) tested the production-perception link in three critical tonal cues—slope, curvature, and turning-point location—and two non-critical cues—mean F0 height and onset F0 height—while Mandarin listeners rated different exemplars of Tone 2. They found that statistically significant correlation was found only for critical cues. In terms of how F0 contour might bias the identification of a segment, Lehnert-LeHouillier (2007) examined German, Japanese, Spanish, and Thai listeners' identification of vowel length, using vowel continua varying orthogonally in both duration (from around 220 ms to 400 ms with a step size of about 30 ms) and F0 contour (level at 180 Hz and falling from 160 Hz to 80 Hz). She found that only Japanese listeners perceived the vowels with a falling F0 as longer; the F0 contour did not seem to have an effect for listeners of other languages. Fogerty and Humes (2012) investigated the contribution of F0, speech envelope, and temporal fine structure in consonants or vowels to overall word and sentence intelligibility. They observed that when dynamic F0 cues were flatted or removed, English listeners still obtained higher recognition scores for vowel-only (i.e., consonantal portions were masked) sentences, as compared to consonant-only (i.e., vocalic portions were masked) ones. These results suggest that dynamic F0 contour might play an important role in consonant identification. However, to the best of my knowledge, no study has systematically investigated how F0 contour alone (e.g., different F0 directions with the same onset F0 height) modulates the perception of voicing of the initial obstruent. While the current study does not set out to examine the respective contribution of post-stop F0 height and F0 contour to the perception of voicing, the potential influence of F0 contour will be addressed in Section 5.4.

2.3. Post-stop F0 at L1 production-perception interface

While there is clear evidence that post-stop F0 functions as a cue for voicing in production as well as in perception *separately*, outcomes from attempts to link the cue use *across* the two modalities remain inconclusive. More generally, based on the proposal that perceptual cue weights arise from statistical regularities in the put (e.g., Holt and Lotto, 2006; Francis et al., 2008; Toscano and McMurray, 2010), one would anticipate the relative informativeness of a cue in a speaker's productions of a contrast to be predictive of the reliance assigned to that cue in perceiving the same contrast. Theories that posit a strong and/or direct connect between production and perception, such as Motor Theory (Liberman and Mattingly, 1985) or Direct Realism (Fowler, 1986), also express such a view. However, although it is established that distributional patterns in production are exploited as cues in perception at the macro level, efforts to find correlations between use of the same cue across production and perception at the micro or individual level have been met with mixed success. For example, while Zellou (2017) found that individuals' production of anticipatory nasal coarticulation on vowels in English was correlated with their patterns of perceptual compensation, Kataoka (2011) found no significant correlation between Californians' production and perception of /u/-fronting in alveolar contexts. Zooming in on the use of post-stop F0, even as the use of post-stop F0 as a perceptual cue for stop voicing reflects the differential F0 at vowel onset in production on a population level, correlational analysis on an individual level has yet to reveal a more direct connection. For instance, the importance an English speaker assigns to post-stop F0 in production does not seem to predict the perceptual reliance of the same cue from the same individual (Shultz et al., 2012). A similar lack of relationship in post-stop F0 cue use for Spanish speakers was reported in Schertz et al. (2020). This study revisits this topic and explores whether there is a direct link between production and perception for the use of post-stop F0 in Mandarin, at both the population and individual levels.

2.4. Post-stop F0 at L2 production-perception interface

If producing and perceiving a phonological contrast means navigating between various acoustic dimensions, learning a

phonological contrast in an L2 then involves adapting the weight associated with relevant dimension to approach that of native speakers of the L2 in question. The majority of work on L2 sound production and perception has put an emphasis on how L2 learners acquire foreign contrasts that rely primarily on dimensions that are not use in similar native contrasts. For instance, the difficulty for Japanese speakers to distinguish the English /r/-/l/ contrast is ascribed to the fact that this English contrast relies mainly on a difference in third formant values, whereas it is the second formant that Japanese speakers use to distinguish the categories (Miyawaki et al., 1975; Iverson et al., 2003; Lotto et al., 2004).

Another interesting line of research focuses on cases in which a first language (L1) contrast primarily relies on *more* cues than the corresponding L2 contrast. A study in this direction is Schertz et al.'s (2015) research on how L1 speakers of Korean, which uses both VOT and post-stop F0 as primary cues for its three-way stop distinction, produce and perceive the L2 English stop contrast, which relies primarily only on VOT.

The current work represents a study that is in some sense sandwiched between the two threads of research discussed above. In particular, similar to English, Mandarin relies primarily on VOT to signal its stop voicing contrast; this therefore distinguishes the case of L1 Mandarin speakers learning the L2 English stop contrast from that of L1 Japanese speakers coping with the English /r/-/l/ contrast. However, this study also deviates from Schertz et al.'s (2015) study of L1 Korean speakers in that, unlike Korean, which uses *both* VOT and F0 as primary cues for its three-way stop contrast, Mandarin only uses F0 as a secondary cue for its two-way stop contrast, but as the primary cue for its lexical tones. Crucially, for L1 speakers of a tonal language learning a non-tonal L2, F0 is an ambiguous cue that signals both tonal and non-tonal (e.g., stop voicing) contrasts in L1, but only non-tonal contrasts in L2. Examining this sort of scenario is therefore important for understanding to what extent L2 learners learn to reweight cues across phonological domains (i.e., using F0 as a dual segmental and suprasegmental cue to using it solely as a segmental cue) during L2 sound category acquisition.

In fact, the research questions raised here have been partially addressed by Guo (2020). In her study, she had a group of Mandarin-English bilinguals dominant in Mandarin produce a set of Mandarin and English words typifying stop voicings in the respective languages, and the same group of participants also took part in 2AFC perception experiments, identifying Mandarin and English words with different combinations of VOT and post-stop F0 values. Visual inspection of her production results suggests that the difference in post-stop F0 between long-lag stops and short-lag stops is smaller in Mandarin than in English, though no statistical models were used to test this observation. In perception, her results also suggest that Mandarin listeners use post-stop F0 as a cue for stop voicing in both L1 Mandarin and L2 English word identification tasks, but whether the extent with which they relied on post-stop F0 differed according to the language context was not analyzed. In this study, these caveats were addressed with a different experiment design.

Much like the link between production and perception in L1, the production-perception interface in L2 has turned out to be elusive, potentially due to more individual variability induced by more diverse L2 learning experiences. While at the broad level, the perception patterns often mirror production patterns, and vice versa, work looking for production-perception links with respect to individual cue weights has had limited luck finding correlation between the two modalities. For example, in studying L1 Korean learners' production and perception of the stop voicing contrast in English, Schertz et al. (2015) find considerable individual difference in L2 English perceptual categorization strategies in spite of the relative homogeneity of their L2 English production. In the current work, the L2 production-perception interface was also briefly examined, focusing on the use of post-stop F0 in L1 Mandarin learners' production and perception of English stops.

2.5. L1 influence on L2 cue use

Given that the target population in this study is L1 Mandarin-L2 English speakers, one would expect the usage patterns of multiple acoustic dimensions in their L2 English to be influenced by their L1 Mandarin. Such an L1-to-L2 influence can be understood in the frameworks of two major theories of L2 speech sound acquisition—the Speech Learning Model (SLM, Flege, 1995, 2007) and the Perceptual Assimilation Model's extension to L2 acquisition (PAM-L2, Best and Tyler, 2007). Both models relate the patterns of L2 sound acquisition to L1 phonology by assuming that L2 sounds are assimilated to L1 sound categories whenever possible. The difficulty of L2 sound discriminability is therefore projected from the phonetic similarity between L1 and L2 sounds, and the patterns of assimilation from L2 to L1 categories. Given that both the English and Mandarin stop contrasts make use of VOT as the primary cue, that the absence/presence of aspiration is an important indicator for phonological voicing, and that both languages have two stop categories in terms of phonological voicing, English phonemically voiced (/b, d, g/) and voiceless (/p, t, k/) stops in the word-initial position will almost certainly be assimilated to Mandarin unaspirated (/p, t, k/) and aspirated stops (/p^h, t^h, k^h/), respectively. In the extreme case where English stops are processed as Mandarin stops, one would expect the participants to transfer their native Mandarin cue-weighting strategies to English, in both production and perception.

However, more recent works have also demonstrated that late L2 learners are able to fine-tune the use of various acoustic dimensions in different language contexts. For instance, Amengual (2021) examined the VOT of the English, Japanese,

and Spanish /k/ in the productions of L1 English-L2 Japanese bilinguals, L1 Japanese-L2 English bilinguals, and L1 Spanish-L2 English-L3 Japanese trilingual and found that all three groups of speakers produced language-specific VOT patterns for each language, despite evidence of cross-linguistic influence. In perception, Casillas and Simonet (2018) investigated whether English beginner learners of Spanish at the early stages of their development could manifest the double phonemic boundary effect in VOT—that is, whether these bilinguals shift the perceptual VOT boundary according to the language mode they are in—and found that they were indeed able to manifest the effect, suggesting that the ability of switching between language-specific perceptual modes can be acquired later in life. It is therefore possible that the bilingual participants in this study are capable of adjusting the weight of post-stop F0 according to the language context. The production and perception experiments presented in this work allow for robust investigation of this possibility.

2.6. Goals of the current study

The use of F0 as a medium for the lexical tones in Mandarin provides an opportunity to examine whether F0 also functions as a cue for stop voicing in production—as has been found for a number of non-tonal languages—and as a cue for stop voicing in perception when Mandarin listeners also need to extract tonal information from F0. With respect to production, previous work has not converged to a definite conclusion, so the current study aims to first establish the post-stop F0 production patterns in the participating speakers. Concerning perception, while there is evidence that Mandarin listeners take advantage of post-stop F0 as a cue for stop voicing, the experiment with which this observation was made did not require the listeners to simultaneously track F0 for lexical tone, so it is therefore still an open question whether Mandarin listeners actually use post-stop F0 as a cue for stop voicing in more natural settings.

The second aim of this study is to investigate whether the use of post-stop F0 cue is sensitive to different language contexts. Capitalizing on the fact that the L1 Mandarin speakers that could be recruited in the university communities here were also L2 English speakers, one relevant question is whether Mandarin-English bilinguals use post-stop F0 cue to different extents, depending on the language "mode" they are operating in. If post-stop F0 is not solely due to physiological and/or aerodynamic reasons and is partially subject to active controlling, as postulated in Kingston and Diehl (1994), Mandarin-English bilinguals might actively, though subconsciously, suppress post-stop F0 in Mandarin because of the pressure to maintain tonal contours, which they do not have to do when speaking English. In perception, the demand to track F0 for lexical tone when perceiving Mandarin might prompt the bilingual listener to attribute variation in F0 partially to lexical tone, which makes them less likely to treat variation in post-stop F0 as an indicator for voicing. However, freed from the burden of tracking F0 for tone, as when they are perceiving English, the same listeners now have more certainty in linking the difference in post-stop F0 to consonantal voicing. These two scenarios could lead to bilinguals using the post-stop F0 cue differentially in both production and perception, which would be reflected as different cue weights for post-stop F0 that depend on the language. On the other hand, given that the bilinguals are dominant in Mandarin, they may simply import their cue-weighting strategies for Mandarin to English, as predicted by the SLM and PAM-L2, resulting in the same weight for post-stop F0, regardless of language. The hypotheses and the corresponding predicted results just described are summarized in Table 1. The conducted production and perception experiments can help distinguish between the two possibilities.

An additional aspect that is foregrounded in this study is individual variability in participants' production and perception in their L1 and L2. Specifically, the relationship between individual participants' production and perception of post-stop F0 is explored. For this purpose, individual participants' production and perceptual post-stop F0 weights in their L1

TABLE 1 Predicted production and perception results under difference hypotheses.

Production	
Hypothesis	**Predicted production results**
Post-stop F0 purely due to physiological / aerodynamic reasons (e.g., Ladefoged, 1967; Ohala and Ohala, 1972; Kohler, 1984) or total transfer of post-stop F0 cue use in Mandarin to English, as prediced by the SLM and PAM-L2	Post-stop F0 difference the same in Mandarin and English tokens
Post-stop F0 partially subject to active controlling (Kingston and Diehl, 1994)	The extent of post-stop F0 difference might depend on the language (i.e., larger in English than in Mandarin)
Perception	
Hypothesis	**Predicted perception results**
Transfer of the Mandarin cue-weighting strategy to English, as predicted by the SLM and PAM-L2	Post-stop F0 weights the same across Mandarin and English
Flexibility in cue use: attributing variation in post-stop F0 partially to lexical tone and partially to stop voicing in Mandarin, but only to stop voicing in English	Post-stop F0 weights depend on the language context (i.e., a higher weight in English than in Mandarin)

and L2 are derived first. Correlation analyses are then used to examine whether individuals' post-stop F0 weights are statistically linked either within the same modality but across languages, or within the same language but across modalities.

3. Production experiment

This experiment examined non-early Mandarin-English bilinguals' productions of Mandarin and English word-initial stops and sonorants on vowel-onset F0.

3.1. Participants

All participants were recruited from the linguistic participant pools at the University of British Columbia or the University of Toronto, and they received partial course credit for participation. A total of 103 participants completed the experiment, but only a subset of 25 L1 Mandarin-L2 English bilingual participants (14 female, 11 male; $\text{Mean}_{\text{age}} = 20.9$ years, $\text{SD}_{\text{age}} = 2.1$ years) were analyzed. The inclusion criteria are detailed below. For their production data to be considered in the analyses, a participant must satisfy all of the following criteria:

1. They completed all required experiment components;
2. They self-report as a native speaker of Mandarin;
3. They have at least one primary caretaker whose native language is Mandarin;
4. They are not simultaneous/early/childhood bilingual in Mandarin and English (i.e., they were exposed to English only after entering elementary school and did not receive their formal education in English prior to high school or university);
5. They lived in China for at least 10 years between birth and age 15.

A number of additional inclusion guidelines, which are based on their audio recording quality and their performance in the perception experiment, were applied to make sure that only high-quality data was included in the analyses. These detailed inclusion guidelines are given in Sections 3.6 and 4.4, respectively. As a preview of these additional criteria, three participants were excluded due to suboptimal recording quality, and only the data from the participants who were attentive throughout the perception experiment was included.

3.2. Stimuli

This section describes the principles behind the selection of Mandarin and English production stimuli. The same logic was used for both languages, with adaptations to accommodate the phonotactic constraints of each language.

3.2.1. Mandarin stimuli

The Mandarin stimuli consisted of 27 monosyllabic Mandarin words in isolation, as provided in Supplementary Table 1. These words had onsets that exemplified the two laryngeal categories—voiceless aspirated and voiceless unaspirated—in Mandarin, as well as the sonorants /m/, /n/, and /l/. The sonorants were included to serve as the baseline against which the phonological voicing of stops was compared. To increase the generalizability of the findings, words with stops at three places of articulation (i.e., labial, alveolar, and velar), crossed with two levels of vowel heights (high: /i/, low: /a/, embedded in /aɪ/; /aɪ/, as opposed to /a/, was used because words with /aɪ/ are phonetically more similar to the English words used in the English production counterpart; see Section 3.2.2), were included. Given that lexical tone has been reported to modulate F0 perturbation in Mandarin (Guo, 2020), and that the influence of individual lexical tones is outside the scope of the current study, only Tone 1 and Tone 4 syllables were considered. Both tones start with a high pitch register and have been found to pattern together in conditioning post-stop F0 perturbation, making their production data more comparable to each other. Note also the existence of systematic and accidental gaps that prevented a fully crossed combination of the onsets, vowels, and tones. For instance, Mandarin disallows the occurrence of a velar stop before a high front vowel, so syllables such as */k^hi/ and */ki/ are missing in Mandarin altogether. It is, however, accidental gaps in the language that cause */maɪ˥/, */ni˥/, etc., to be absent.

The stimuli were presented to the participants in simplified Chinese characters. Given that Mandarin has a large number of homophones that are nonetheless distinguished by different characters, each stimulus was represented with a common character so that all of them should be familiar to the participants, with the exception of *kai4* 忾, which is not a highly frequent character. To make sure that the participant knew the pronunciation of this character, its pinyin <kai4> was added to the right side of this character when presented to the participant. Care was also taken to ensure that different characters were as visually distinct as possible, to avoid the potential confound from visual priming across trials. For instance, while *pi1* could be represented with both 披 and 批, 披 was chosen because 批 shares the component 比 with another stimulus *pi4* 屁.

3.2.2. English stimuli

The English stimuli consisted of 19 monosyllabic words, as given in Supplementary Table 2. These words were selected following the same principles of stimulus section for the Mandarin tokens: the onsets typified voiceless stops, voiced stops, and sonorant at labial, alveolar, and velar places, while the vowels were either the front high vowel /i/ or the diphthong /aɪ/. When a simple combination of an onset and an open vowel did not correspond to a common English word, another common

word with the same onset and nucleus but with an additional voiceless-stop coda was used as the alternative. Voiceless-stop codas, instead of other consonant classes, were used because they formed common English words. Also, for the syllable /di/, both the letter *D* and the word *deep* were used as stimuli to prevent loss of data for /di/ due to the participant not producing /di/ upon seeing *D*.

3.3. Procedure

The procedure was identical for both the Mandarin and English versions of the experiment, and the order in which the two versions were administered was counterbalanced across participants. The entire experiment took place online in response to constraints on in-person data collection due to COVID, with the participant being instructed to complete the experiment on their own computer in a quiet place. They were encouraged to use an external microphone to keep the fidelity of audio recordings as high as possible, though they could still participate using the built-in microphone on their device.

The experiment was implemented in jsPsych, version 6.1.0 (de Leeuw, 2015). The experiment started with a microphone check to ensure that the input source was set correctly, and that the recording was clear. The experimental trials commenced after three practice trials that aimed to familiarize the participant with the recording interface and experimental flow. Each stimulus was repeated three times in three blocks, respectively with a self-timed break between blocks. Stimuli were presented in a randomized order within each block. Each trial began with a plus sign at the center for 500 ms, and the recording was initiated automatically at the same time. The stimulus then appeared at the center, replacing the plus sign, and the participant was asked to read aloud the stimulus in a clear and natural manner. The trial ended with the participant clicking the "submit" button, which stopped the recording, uploaded the audio file to the server, and triggered the next trial. In the event where the participant did not click anything, the trial would terminate on its own after 10 s. The entire production experiment lasted about 15 min.

3.4. Recording annotations

All annotations and measurements were performed in Praat (Boersma and Weenink, 2021). The portion of the signal analyzed spanned from the beginning of the onset consonant to the end of the third pitch cycle of the nucleus vowel. The following guidelines were used when annotating tokens produced in either language.

1. *Beginning of stop closure voicing*: In the cases where there was prevoicing for tokens with a voiced stop in English or, very rarely, with an unaspirated stop in Mandarin, all simple periodic chunks of the waveform before the release of the onset stop were marked as stop closure voicing.
2. *Beginning of stop burst*: For tokens with a stop onset, the beginning of the burst was marked at the starting point of perturbation in the waveform.
3. *Vowel onset*: The vowel onset was operationalized as the point where the (quasi) periodic part of the vowel first crossed zero in the positive direction.
4. *End of the third pitch cycle*: Following Cole et al. (2007) and Clayards (2018), the point marking the first 3 pitch cycles as counted from vowel onset was pinned in order to derive the onset F0.

3.5. Acoustic measurements

1. *Voice Onset Time (VOT)*: In line with the typical definition, VOT is defined as the time difference between the release of the stop and the onset of voicing (pre- or post-release). Accordingly, for prevoiced tokens (i.e., those with the beginning of stop closure voicing marked) VOT took a negative value, while VOT was positive for tokens where the onset of vocalic voicing followed the stop release. Tokens where the onset of vocalic voicing coincided with the stop release had a VOT of 0 ms.
2. *Onset fundamental frequency (F0)*: This measurement was obtained by dividing 3 by the duration of the first 3 pitch cycles from vowel onset [i.e., 3 / (end of the third pitch cycle − vowel onset)]. No F0-tracking algorithm was therefore involved for this measurement.

3.6. Participant inclusion criteria

Participants whose entire recordings (i) contained excessive background noise due to their doing the experiment in a noisy place ($n = 1$), (ii) were extremely soft that made it challenging to identify acoustic landmarks for annotation ($n = 1$), or (iii) were of extremely low sampling rates ($n = 1$), were omitted from the dataset altogether. There were also three participants who attempted the experiment more than once; in such a case, only the recordings from their first experiment attempt were considered. A subset of 25 participants was then selected based on their performance in the perception experiment, as explained in Section 4.4.

3.7. Omitted data

Among the tokens produced by the 25 included participants, the following tokens were excluded from all analyses: mispronunciations (11 Mandarin and 26 English), skipped

tokens (2 Mandarin and 3 English), and technical issues (2 Mandarin and 4 English, including sporadic silent periods that overlapped with stop burst and/or vowel onset). Furthermore, tokens with creaky voice at vowel onset, for which F0 estimation was therefore unreliable, were also omitted from all analyses (50 Mandarin and 33 English). Overall, 131/3,450 = 3.8% of the production tokens were excluded.

3.8. Statistical analyses

The analyses consisted of two major parts: the first part addressed whether post-stop F0 had different values across the onset types in each language, and the second part focused on the quantification of production weight for post-stop F0 in each language. All models were fitted with Bayesian mixed-effects models, using `CmdStanR` (Gabry and Češnovar, 2021), an R interface for the Stan probabilistic programming languages (Carpenter et al., 2017). Bayesian models were chosen because they return a distribution of potential values for all model parameters, making it more intuitive to assess the uncertainty associated with each parameter. In what follows, details about the statistical model employed are described.

3.8.1. Post-stop F0 models

In this set of analyses, post-stop F0 was modeled as a Gaussian linear function of a number of variables that were properties of tokens or speakers. The names of predictor variables are given **boldface**, and different levels within a variable are indicated in SMALL CAPS.

3.8.1.1. Variables

The dependent variable in all models was z-transformed post-stop F0. The post-stop F0 values from both Mandarin and English production were z-transformed within each speaker. That is, a single z-transformation was applied to Mandarin and English production data together for each speaker.

Four token-level predictors were considered: the **voicing** of the onset consonant, **language/tone**, the **height** of the main vowel, and the **place of articulation** (**PoA**) of the onset consonant. Forward difference coding was used for **voicing** (ASPIRATED vs. UNASPIRATED and UNASPIRATED vs. SONORANT). Helmert coding was used for **language/tone** (ENG vs. mean of MAN T1 and MAN T4, and MAN T1 vs. MAN T4). Sum coding was used for **height** (HIGH, NON-HIGH = [1, −1]) and **PoA** (LABIAL, ALVEOLAR, VELAR, with LABIAL coded with −1). To account for how each predictor affected the realization of the voicing contrast, two-way interaction terms between **voicing** and all the other predictors were also included in the model comparison process. These first-order and second-order terms therefore constituted the population-level ("fixed-effect") predictors.

For individual-level ("random-effect") predictors, by-**speaker** effects consisted of a random intercept and random slopes for all population-level predictors.

3.8.1.2. Model structure

Standardized post-stop F0 was modeled as a function of a subset of the predictor variables introduced above, using Bayesian linear mixed-effects models. All candidate models shared general specifications. Main-effect terms were included for the predictor variables selected in a particular candidate model. As mentioned above, two-way interaction terms being **voicing** and the other predictors were also considered. I did not, however, consider any three-way interactions as they are in general harder to interpret and could drastically slow down model sampling. All models also included by-speaker random intercepts, to account for variability in post-stop F0 of speakers beyond the effects of predictor variables. All possible by-speaker random slopes were also included to account for variability among speakers in the effects of predictors on post-stop F0 (Barr et al., 2013).

Each model was fitted with regularizing priors of Normal($\mu = 0, \sigma = 5$) for the intercept and all population-level parameters. An Exponential($r = 1$) distribution was used as the prior for the error term as well as for the individual-level standard deviations. Correlations among individual-level effects used the LKJ prior (Lewandowski et al., 2009) with $\xi = 1$, in order to give lower prior probability to perfect correlations. All models showed no divergent transitions and had $\hat{R}$ values close to 1 (i.e., all $\hat{R} < 1.01$), which indicates that chains were well-mixed.

3.8.1.3. Inference criteria

Evidence embedded in each model was evaluated in two ways: (i) the posterior distributions of parameters, and (ii) comparison of models of different complexities. In particular, I consider there to be strong evidence for a non-null effect if the 89% credible interval (CrI)—the narrowest interval that contains 89% of the posterior density—for the parameter does not include 0. If the 89% CrI spans 0, but the probability of the parameter not changing direction is at least 89%, I consider this to represent weak evidence for a given effect. The decision to use CrIs of 89%, as opposed to 95%, is based on Koster and McElreath (2017) and McElreath (2020), to discourage the association between a Bayesian posterior distribution and a p-value. Model comparison was done by means of the Bayesian leave-one-out estimate of expected log pointwise predictive density (ELPD-LOO; Vehtari et al., 2017), which aims to gauge a model's *predictive* accuracy (i.e., how close predicted values from a model are to the raw data). A higher ELPD-LOO value means that the model has a better predictive accuracy. The results from model comparison thus inform us whether a variable contributes substantially to a model's predictive power. Following Sivula et al. (2020), when the estimated absolute difference in ELPD-LOO between two models is at least 4, and 0 is not within two

TABLE 2 Candidate post-stop F0 models considered in model comparison, with their ELPD-LOO means and standard errors.

Model	ELPD-LOO mean	ELPD-LOO standard error	Predictors
M1	−3637.3	60.3	height + lang/tone
M2	−3221.8	67.3	height + lang/tone + voi
M3	−3215.5	67.3	height + lang/tone + voi + PoA
M4	−3205.5	68.2	height + lang/tone + voi + voi × height
M5	−3189.0	67.8	height + lang/tone + voi + voi × lang/tone
M6 (final)	−3173.4	68.7	height + lang/tone + voi + voi × height + voi × lang/tone
M7	−3174.3	69.2	height + lang/tone + voi + voi × height + voi × lang/tone + voi × height × lang/tone

An intercept was included in each model but is omitted here in the table to save space.

standard errors of the estimated difference, there is evidence that the two models give different predictions.

In the following sections, model parameters are reported in terms of marginal posterior means of parameters, 89% CrIs, and the probability of effect direction.

3.8.1.4. Candidate models

The construction of candidate models for model comparison relied both on prior knowledge about factors affecting post-stop F0 and on a compromise between model complexity and predictive accuracy. All the candidate models are given in Table 2. Given that vowel height is known to influence F0 ("intrinsic F0," Whalen and Levitt, 1995) and that language and lexical tone can affect F0, the base model (i.e., M1) started with the factors **height** and **language/tone**. As one of the goals is to establish whether and how post-stop F0 might be influenced by phonological voicing, further models were constructed by incrementally adding terms that involved **voicing**. For example, the comparison between M1 and M2 assessed the contribution of voicing in predictive accuracy, and comparing M2 and M4 examined the importance of the interaction between voicing and vowel height in predicting post-stop F0 values. Furthermore, a model with **PoA** as a predictor (i.e., M3) also entered into comparison to confirm that place of articulation does not cause post-stop F0 to differ. The formal specification of the final model can be found in the Supplementary material.

3.8.2. Post-stop F0 production weight model

The second set of analyses aimed to quantify the production weight associated with post-stop F0. A higher production weight means post-stop F0 is more reliable in separating different members of the contrast. Following Clayards (2018), the production weight was calculated based on the amount of overlap between the categories, which was quantified using Cohen's *d* (Cohen, 1988):

$$d = \frac{\mu_{\text{asp}} - \mu_{\text{unasp}}}{\sqrt{1/2\left(\sigma^2_{\text{asp}} + \sigma^2_{\text{unasp}}\right)}},$$

where μ_{asp} and μ_{unasp} refer to the mean F0s of the aspirated and unaspirated categories, respectively, and σ^2_{asp} and σ^2_{unasp} are the standard deviations of F0 of the aspirated and unaspirated categories, respectively.

Cohen's *d* for post-stop F0 was calculated at the population level with all speakers as a whole and at the individual level for each speaker. Only tokens produced with a positive VOT were included in the calculation, as negative VOTs were rare in the data (i.e., 9 tokens from 1 speaker in Mandarin, and 40 tokens from 5 speakers in English) and therefore were not representative of the norm of this speaker population. Additionally, rather than estimating cue weights from empirical data as in most previous work (e.g., Shultz et al., 2012; Schertz et al., 2015; Clayards, 2018), a statistical model was used to derived the weight, which allowed for uncertainty around the weight to be incorporated. For this purpose, a Bayesian mixed model was first fitted to obtain the means and standard deviations of F0 of the aspirated and unaspirated categories for the whole group and for each speaker. The model included a cross-category correlation structure and used partial pooling to estimate individual means and standard deviations. For instance, a speaker's mean post-stop F0 for the aspirated category was correlated with their mean post-stop F0 for the unaspirated category, and both mean values were informed not only by the speaker's own production data, but also by other speakers' data thanks to partial pooling. The estimated means and standard deviations were then fed to the Cohen's *d* formula above to calculate the production weight within the model. As such, the post-stop F0 weights of the entire group and for each speaker were not just a single numerical value but a *distribution* that also carried information about uncertainty. The formal specification of the model is included in the Supplementary material.

3.9. Results: Production of post-stop F0

Mean production values and standard deviations for L1 Mandarin and L2 English stops and sonorants on VOT and post-stop F0 are given in Supplementary Table 3. Distributions

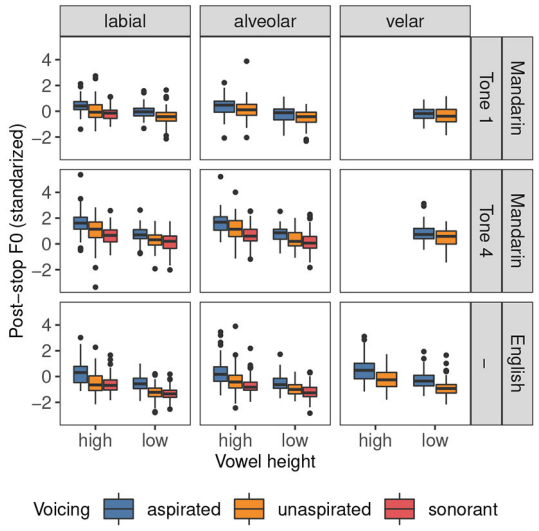

FIGURE 1
Standardized post-stop F0 values, normed by speaker, as a function of place of articulation, language/tone, vowel height, and phonological voicing.

of standardized post-stop F0 values are plotted in Figure 1. ELPD-LOO means and standard errors for the candidate models are listed in Table 2. A higher ELPD-LOO value means the model has a better predictive accuracy, so, for example, M2 makes better predictions than M1. Finally, model comparison results are summarized in Supplementary Table 4 in terms of difference in ELPD-LOO values and associated standard errors. Note that the difference score in each cell was computed by subtracting the ELPD-LOO value of the model represented in the column from the ELPD-LOO value of the model indicated in the row. For instance, the difference −415.5 came from $\text{ELPD-LOO}_{M1} - \text{ELPD-LOO}_{M2} = (-3637.3) - (-3221.8)$.

The results of model comparison indeed confirmed the importance of phonological voicing in conditioning post-stop F0 (i.e., M1 vs. M2) and spoke to the importance of interaction between voicing and vowel height (i.e., M2 vs. M4), and between voicing and language/tone (i.e., M2 vs. M5). Place of articulation, however, did not seem to influence post-stop F0 (i.e., M2 vs. M3). Since no significant gain in prediction was observed past M6, M6 was selected as the best balance between model complexity and predictive performance among the models being compared. The interpretation and discussion presented below are therefore based on this model.

In presenting the results, summary statistics and visualizations derived from raw data are given first, followed by the output from the final model in terms of posterior distributions for key parameters. I first interpret population-level parameter estimates before moving on to individual-level estimates.

3.9.1. Population results

The marginal posterior distributions for population-level parameters from M6 are summarized in Table 3. As expected, both vowel height and language/tone contribute to difference in post-stop F0. Specifically, the high vowel /i/ led to a higher onset F0 (HIGH − mean height: $\bar{\beta} = 0.32$, 89% CrI = [0.27, 0.36], $p(\beta > 0) = 1.00$), and Tone 4 tended to have a higher onset F0 than Tone 1 (MAN T1 − MAN T4: $\bar{\beta} = -0.91$, 89% CrI = [−1.03, −0.79], $p(\beta < 0) = 1.00$). Also, participants' L2 English tended to have a lower onset F0, in comparison with their L1 Mandarin (ENG − (MAN T1 + MAN T4)/2: $\bar{\beta} = -0.84$, 89% CrI = [−1.02, −0.66], $p(\beta < 0) = 1.00$), which agrees with the general finding from the literature (Keating and Kuo, 2012; Lee and Sidtis, 2017). Critically, in both languages, aspirated stops had a higher post-stop F0 than unaspirated stops (ASP − UNASP: $\bar{\beta} = 0.49$, 89% CrI = [0.41, 0.56], $p(\beta > 0) = 1.00$), which in turn had a higher post-stop F0 than sonorants (UNASP − SON: $\bar{\beta} = 0.29$, 89% CrI = [0.20, 0.39], $p(\beta > 0) = 1.00$). In addition, the extent of post-stop F0 difference due to aspiration was contingent on language and tone as well, such that bilingual speakers' English tokens showed an even bigger difference than Mandarin tokens ([ASP − UNASP] × [ENG − (MAN T1 + MAN T4)/2]: $\bar{\beta} = 0.25$, 89% CrI = [0.10, 0.39], $p(\beta > 0) = 1.00$), and so did their Mandarin Tone 4 tokens in comparison with Tone 1 tokens ([ASP − UNASP] × [MAN T1 − MAN T4]: $\bar{\beta} = -0.16$, 89% CrI = [−0.28, −0.05], $p(\beta < 0) = 0.99$).

3.9.2. Individual results

The distributions for key parameters involving voicing for each participant are visualized in Figure 2. In both their Mandarin and English productions, there is strong evidence that all speakers produced a higher post-stop F0 following an aspirated stop than an unasiprated stop, as the 89% CrI is above 0 for all speakers in the [ASP − UNASP] panel in Figure 2. The [UNASP − SON] panel indicates that, for the majority of speakers (18 out of 25), the model is also confident that their onset F0 was higher adjacent to an unaspirated stop than adjacent to a sonorant. For the remaining speakers, even though their 89% CrIs span 0, their posterior means are still above 0, suggesting that, on average, their F0 patterns conform to the general trend. In terms of the post-stop F0 difference due to aspiration, about half of the speakers (13) evidently agree with the population pattern in having a bigger F0 difference in English, as indicated by their positive 89% CrIs in the [(ASP − UNASP) * LANG] panel. For the other speakers, there does not seem to be a consistent trend, as even the posterior means are going in different directions. Finally, as shown in the [(ASP − UNASP) * TONE] panel, even though only seven speakers clearly followed the observation at the population level that Tone 4 supported a more differentiated post-stop F0 distinction between aspirated and unaspirated stops, the other speakers also trend in this direction.

TABLE 3 Marginal posterior summaries for key population-level parameters from M6.

Parameter	Mean	SD	89% CrI	p(dir.)
intercept	0.01	0.01	[−0.01, 0.04]	$p(\beta > 0) = 0.84$
HIGH − (HIGH + LOW)/2**	0.32	0.03	[0.27, 0.36]	$p(\beta > 0) = 1.00$
ENG − (MAN T1 + MAN T4)/2**	−0.84	0.11	[−1.02, −0.66]	$p(\beta < 0) = 1.00$
MAN T1 − MAN T4**	−0.91	0.07	[−1.03, −0.79]	$p(\beta < 0) = 1.00$
ASP − UNASP**	0.49	0.05	[0.41, 0.56]	$p(\beta > 0) = 1.00$
UNASP − SON**	0.29	0.06	[0.20, 0.39]	$p(\beta > 0) = 1.00$
[ASP − UNASP] × [HIGH − (HIGH + LOW)/2]*	0.05	0.04	[−0.01, 0.12]	$p(\beta > 0) = 0.91$
[UNASP − SON] × [HIGH − (HIGH + LOW)/2]	0.04	0.04	[−0.02, 0.10]	$p(\beta > 0) = 0.86$
[ASP − UNASP] × [ENG − (MAN T1 + MAN T4)/2]**	0.25	0.09	[0.10, 0.39]	$p(\beta > 0) = 1.00$
[ASP − UNASP] × [MAN T1 − MAN T4]**	−0.16	0.07	[−0.28, −0.05]	$p(\beta < 0) = 0.99$
[UNASP − SON] × [ENG − (MAN T1 + MAN T4)/2]	−0.03	0.08	[−0.15, 0.09]	$p(\beta < 0) = 0.64$
[UNASP − SON] × [MAN T1 − MAN T4]	0.00	0.10	[−0.16, 0.15]	$p(\beta < 0) = 0.52$

The contrast coding scheme for each variable is described in Section 3.8. The parameters whose effects are judged to be strong are marked with **, and those whose effects are judged to be weak are marked with *.

3.10. Results: Production weights of post-stop F0

Standardized post-stop F0 values are plotted against raw VOT values for participants' Mandarin and English productions in Supplementary Figure 1, and the distributions of production VOT and post-stop F0 weights, expressed in terms of Cohen's *d*, at the population level are graphed in Figure 3. Although the focus on this study is on the post-stop F0 cue, for completeness, the results for the VOT weight are also reported below.

3.10.1. Population results

As can be seen in Figure 3, speakers as a group had a much higher weight for VOT than for post-stop F0, in both their Mandarin and English production. Also, regardless of language, there was more uncertainty surrounding the post-stop F0 weight than the VOT weight, as measured by the coefficient of variation (CV), which is defined as the ratio of the standard deviation to the mean (English: $CV_{VOT} = 0.06$, $CV_{F0} = 0.18$; Mandarin: $CV_{VOT} = 0.06$, $CV_{F0} = 0.17$). Contrasting the weights along the same dimension across languages, more weight was assigned to VOT in the Mandarin production (89% CrI = [6.34, 7.60]), as compared to the English production (89% CrI = [4.78, 5.82]), while the converse was true for the post-stop F0 weight: English tokens showed a heavier reliance on post-stop F0 (89% CrI = [0.70, 0.99]) than Mandarin tokens (89% CrI = [0.34, 0.54]).

3.10.2. Individual results

The reliability of each dimension for individual speakers, as estimated by Cohen's *d*, is plotted in Figure 4. Conforming to the population pattern, all speakers assigned more weight to VOT than post-stop F0 in both their Mandarin and English productions (Figure 4A). When correlating weights along the two dimensions within language, no specific correlation pattern was discernible (see Figure 4B; Mandarin: 89% CrI of $\rho_{VOT_{Man},F0_{Man}} = [-0.35, 0.15]$; English: 89% CrI of $\rho_{VOT_{Eng},F0_{Eng}} = [-0.25, 0.21]$). However, when the VOT weights were correlated across languages, a strong positive correlation was observed (89% CrI of $\rho_{VOT_{Man},VOT_{Eng}} = [0.40, 0.81]$), indicating that speakers who showed a larger VOT weight in Mandarin also tended to have a larger VOT weight in English (Figure 4C). In addition, for all but one speaker, VOT had more weight in their Mandarin tokens than their English tokens. For the post-stop F0 weight, most individuals (19 out of 25) echoed the population pattern in shifting their F0 weight upward when producing English tokens (Figure 4D), although there was no correlation in this cue across languages (89% CrI of $\rho_{F0_{Man},F0_{Eng}} = [-0.49, 0.40]$). Also notice that there was more individual variation for the post-stop F0 weight in the English production than in the Mandarin production, as indicated by a wider spread of individual weights in English than in Mandarin.

3.11. Interim discussion: Production

The Mandarin production results reported here are in line with the recent work by Guo (2020) in terms of post-stop F0: both at the population and individual levels, the vowel-onset F0 following aspirated stops was higher than that following unaspirated stops. In addition, for most speakers, vowel-onset F0 after unaspirated stops was in turn higher than that after sonorants. Similar to their Mandarin production, the participants' English production also demonstrated a difference in post-stop F0 between aspirated and unaspirated series, but

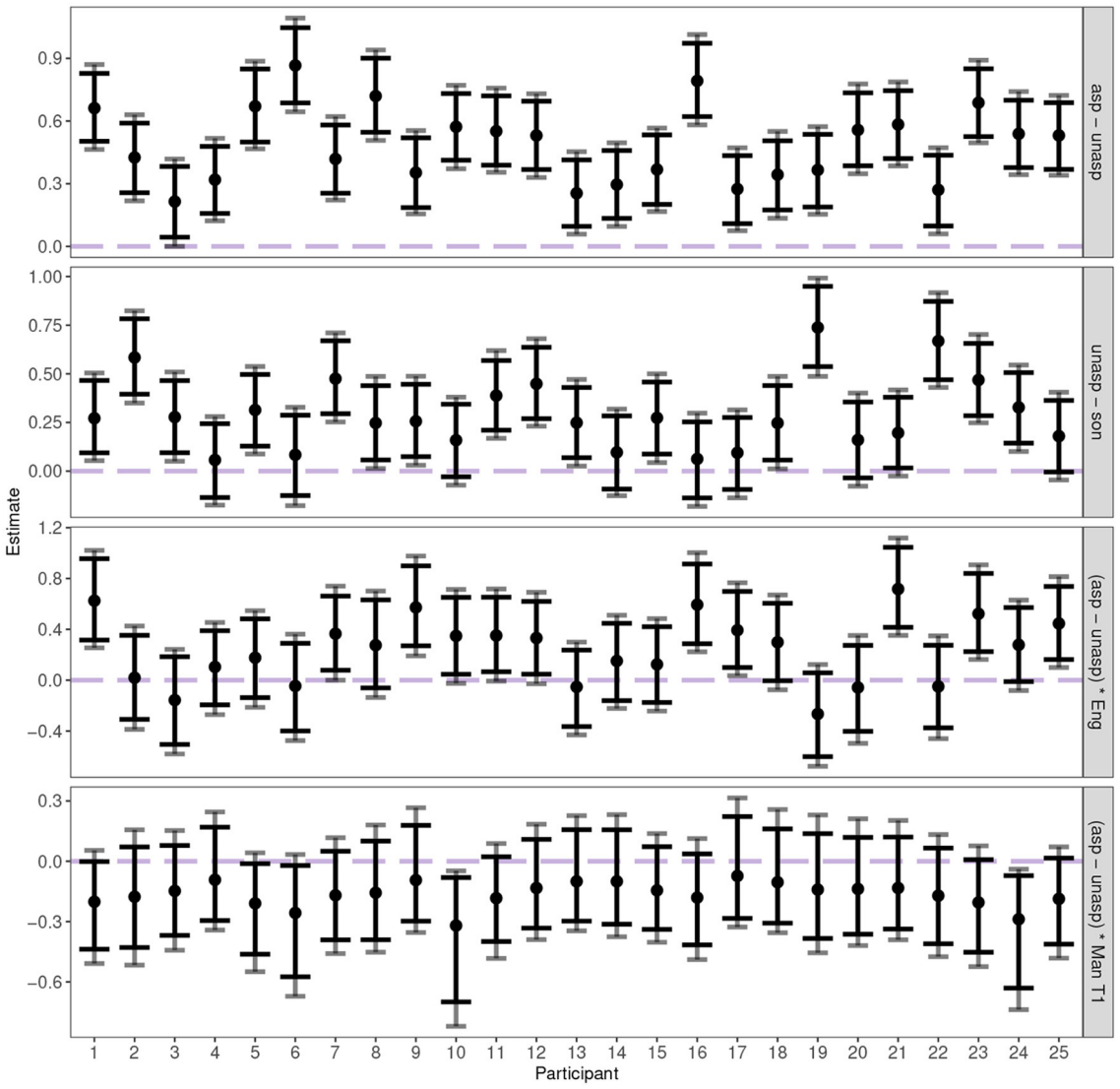

FIGURE 2
Marginal posterior summaries for key parameters involving voicing for each individual speaker. The [asp – unasp] panel shows the difference in F0 between aspirated and unaspirated stops. The [unasp – son] panel shows the difference in F0 between unaspirated stops and sonorants. The [(asp – unasp) * Eng] panel shows the further difference in F0 between aspirated and unaspirated stops in English, in comparison to Mandarin. The [(asp – unasp) * Man T1] panel shows the further difference in F0 between aspirated and unaspirated stops in Mandarin Tone 1 tokens, when compared to Tone 4 tokens. The dots denote the posterior means. The inner error bars represent 89% CrIs, and the outer error bars represent 95% CrIs.

with an even larger F0 gap, both for the speakers as a whole and for over half of the individual speakers. This pattern again agrees with what has been found in Guo (2020).

Regarding cue weighting, VOT was the most reliable dimension distinguishing aspirated from unaspirated stops in both Mandarin and English, though it seemed that VOT assumed an even higher weight in Mandarin for almost all speakers (as measured by the posterior mean). The opposite pattern was observed for the post-stop F0 weight: English induced a higher weighting in this cue for most speakers. When the weighting between the two cues was correlated within each language, however, neither an enhancing nor a trading relationship was obtained.

4. Perception experiment

The perception experiment turns to the perception of the Mandarin and English stop contrasts in the word-initial position by the same L1 Mandarin-L2 English bilinguals. The focus is on the contribution of post-stop F0 to categorization of the contrasts.

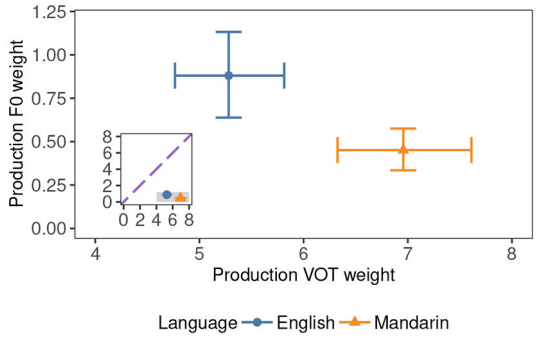

FIGURE 3
Group-level production weights for VOT and post-stop F0 in each language. The embedded plot shows the same data but with the same scale along both axes, and the dashed line is $y = x$, which represents equal production weights for both dimensions. The shaded area indicates the part enlarged in the main plot.

4.1. Participants

The same group of participants from the production experiment also took part in the perception experiment. The perceptual data analyzed here came from the same 25 participants whose production tokens were analyzed in the production experiment.

4.2. Stimuli

All stimuli were created from natural productions of the Mandarin words *bi1*, *pi1*, *bi4*, *pi4*, *yi1*, *mi1*, *mi4*, and *ni4* read by a 24-year-old male English-Mandarin speaker who speakers English as L1 but is also fluent in Mandarin. The prompts for production were words in isolation, which were presented three times to the model speaker in a randomized order. The recording was made on the Sound Devices MixPre-D audio mixer with a headset microphone. The produced syllables were then scrutinized by the author, and one token that was clear and did not have creaky quality was selected for each word as the raw tokens for manipulation.

4.2.1. Mandarin stimuli

Stimuli could be categorized into the target or filler sets, with both sets containing Tone 1 and Tone 4 syllables. The target set was composed of syllables with a bilabial stop as the onset and the high vowel [i] as the nucleus, with the VOT of the stop and the initial F0 contour of the vowel manipulated. The manipulation along the VOT and F0 dimensions is summarized in Figure 5 and explained in the following paragraphs. Bilabial stops were used because they do not have lingual targets and therefore are expected to be coarticulated to a lesser degree with the following vowel (Schertz et al., 2020). The vowel

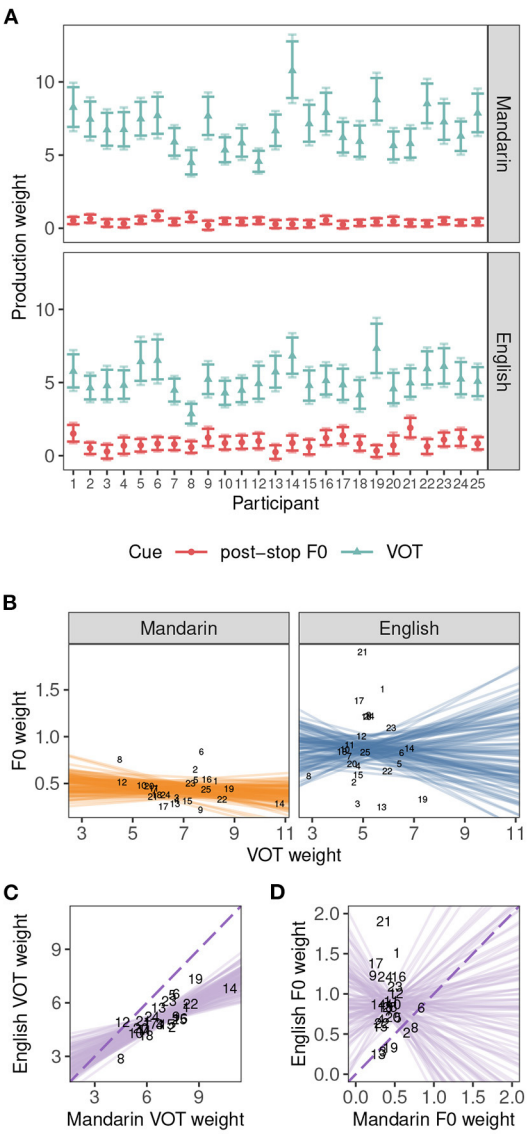

FIGURE 4
(A) Individual speakers' production weights for VOT and post-stop F0. The posterior means are represented by the dots. The 89% CrIs are represented by the inner error bars, and the 95% CrIs are represented by the outer error bars. **(B)** Post-stop F0 weights against VOT weights, separately for each language. **(C)** Production VOT weights across languages. **(D)** Production post-stop F0 weights across languages. In **(B–D)**, the solid lines represent 100 regression lines fitted with 100 posterior draws, to show the direction and uncertainty of the correlation. The dashed line is $y = x$, where the Mandarin weight equals the English weight.

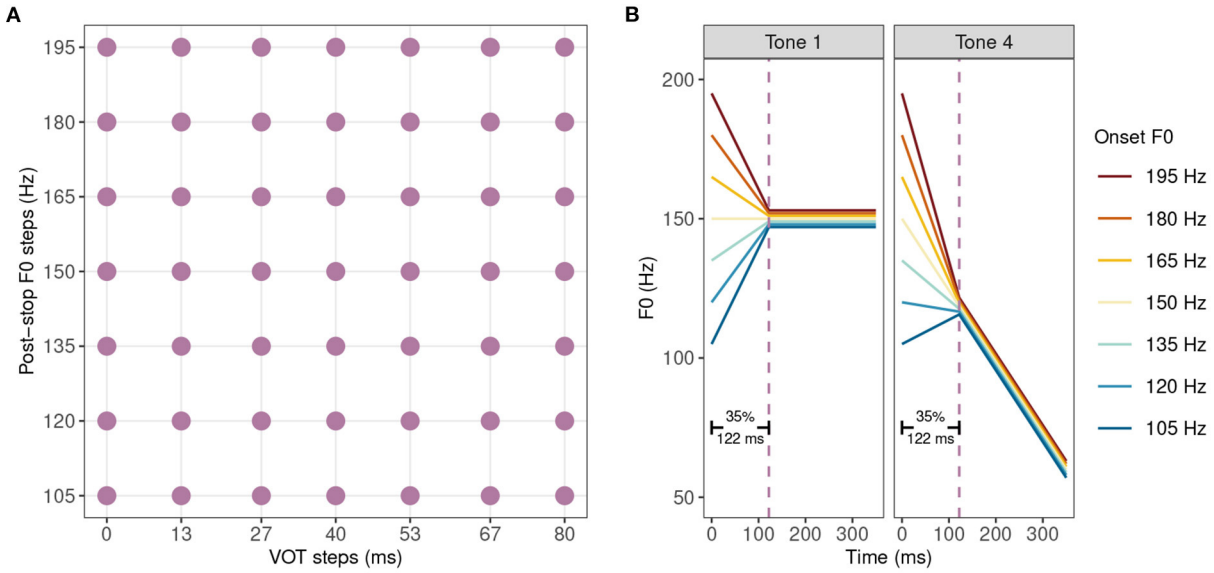

FIGURE 5
Manipulation of target stimuli for all perception experiments. **(A)** Each dot represents one stimulus, with its *x* coordinate corresponding to the VOT of the initial labial stop, and its *y* coordinate to the initial F0 of the following vowel. **(B)** Illustration of F0 trajectory manipulation for target syllables. Note the vowel duration in actual stimuli is not necessarily 350 ms due to the trade-off between VOT and vowel duration, which was also manipulated. The invariant parts across different tokens (i.e., after 122 ms) are shifted vertically in the figure for visual clarity only.

[i] was selected because its formants are more stable across time in general (Hillenbrand et al., 1995)[1]. In addition, the combination of bilabial stops with the high vowel also led to valid English lexical items *pea* and *bee*; this was critical given that the exact same stimuli were used in the English version of the experiment as well. For fillers, Mandarin words *yi1*, *yi4*, *mi1*, and *mi4* were selected because they typified other onset types than the stop.

The target syllables were created by cross-splicing the vocalic portion of the *bi1* token and the burst+aspiration portion of the *pi1* token. The detailed steps of stimulus manipulation are described below.

The first step involved creating a Tone 1 and a Tone 4 base token for downstream manipulation. The vowel duration of the *bi1* token was set to 350 ms, which is approximately the mean duration of 416.2 ms for citation Tone 1 syllables and 307.8 ms for citation Tone 4 syllables (Yang et al., 2017)[2]. The vowel duration was shifted to an ambiguous value to discourage the participant to use it as an additional cue for tone identification (e.g., Blicher et al., 1990). F0 trajectories were then manipulated to mimic natural Tone 1 and Tone 4 contours. For Tone 1, a simple pitch stylization was applied by setting both the initial and final F0 on the vowel to 150 Hz. The F0 was set to 150 Hz because this was very close to the natural Tone 1 F0 register of this particular token. Tone 4 was stylized as a linear F0 decline from 150 Hz to 60 Hz. The initial 150 Hz was to match the initial F0 value for Tone 1 while the final 60 Hz was set based on the model speaker's natural Tone 4 production. The decision to recreate Tone 4 F0 contour from a Tone 1 item, instead of using a natural Tone 4 item, was to make sure that the same intensity profile was shared and would not be a confound[3].

1 The vowel [i] was also preferred from the perspective of VOT manipulation. Given that the starting values of the formant frequencies in the voiced part of the vowel could be substantially different depending on VOT, stimuli whose VOT values are manipulated with a "progressive cutback and replacement" approach (which was also used in this study) can have initial formant frequencies being correlated with VOT, leading formant cues to be a confound. Winn (2020) argues that since F1 of [i] is already low, the upward F1 transition common to the other vowels would be minimized, thus offering no covarying cue for VOT.

2 These measurements are based on production of isolated monosyllables by 121 speakers (46 male and 75 female). Note that even though there seems to be a 100-ms difference between Tone 1 and Tone4, both tones have a standard deviation of about 90 ms in the syllable duration measurement, suggesting that the two tones overlap to a large extent in terms of their duration distributions.

3 I have also attempted to create base tokens in the opposite direction: creating a Tone 1 item from a Tone 4 item. However, the resulting audio was noticeably unnatural, especially in the later portion where F0 needed to be raised from a low target of Tone 4 to a high target of Tone 1.

The second step scaled the intensity of the two base tokens to 75 dB based on the root-mean-square (RMS) amplitude. The level 75 dB was chosen because this was approximately the intensity of the raw recording. Intensity normalization was done at this step, as opposed to at a later point when actual stimuli were synthesized, because Winn (2020) cautions that "the inclusion of a lengthy aspiration portion will justifiably reduce overall RMS intensity, so equalization would result in unnatural amnlification of the syllable with voiceless onset" (p. 859). He therefore suggests that intensity amplification/attenuation should be applied before initiating VOT manipulation.

In the last step, the two intensity-equalized tokens were then modified, using a Praat script prepared by Winn (2020), to create tokens varying in VOT duration and F0 at vowel onset. The duration of VOT in the base tokens was manipulated on a 7-step series ranging from 0 ms to 80 ms. The range endpoints were meant to span the VOTs of both English and Mandarin word-initial bilabial stops while still having enough resolution. Note that negative VOT was not in the manipulated range partially because "voiced" stops in word-initial position in English are very often realized as a short-lag stop with positive VOT (Fulop and Scott, 2021) and partially because including negative values would decrease the manipulation resolution. VOT was manipulated with a progressive-cutback-and-replacement approach—that is, "the onset of a word with a voiced stop sound is progressively deleted and replaced with a roughly equivalent amount of the onset from its voiceless-onset counterpart" (Winn, 2020, p. 854)—to accommodate the observation that there tends to be an inverse relationship between VOT and duration of the following vowel (Summerfield, 1981). However, to approximate this inverse relationship in natural production, the extent of vowel shortening was not entirely commensurate with changes in VOT, that is, for every 1 ms of VOT increase, the vowel was shortened by less than 1 ms (Allen and Miller, 1999; Toscano and McMurray, 2010). The default vowel-VOT ratio of 0.65, which is the default value of Winn's (2020) script, was used for modeling this trade-off relation. The initial F0 was set at one of the seven values, from 105 Hz to 195 Hz with a step size of 15 Hz, at the beginning of the vowel. F0 then rose/fell linearly for the following 122 ms (or 35% of the vowel duration) to 150 Hz for Tone 1 stimuli and to about 118 Hz for Tone 4 stimuli. The step size was set to 15 Hz so that the difference in F0 would be large enough to be noticeable but not too large so as to distort the F0 trajectory significantly, and the temporal extent of manipulation was fixed at 35%, following the practice in Guo (2020), which was in turn based on the Mandarin production data in her study. Note that, as pointed out by one reviewer, the F0 manipulation resulted in initial F0 trajectories that differed not only in onset F0 but also in F0 contour (see Figure 5B). The F0 cue here therefore involved both F0 height and direction.

The creation of filler items roughly followed the same first two steps in creating the target items (e.g., [i˥] and [i˩˧] were created from a natural production of *yi1*), except that the filler [mi˩˧] was modified from a natural Tone 4 syllable, *mi4*, rather than being constructed from the Tone 1 syllable *mi1*. However, the tonal contour of this filler item was similarly styled to that of target Tone 4 items, to prevent this filler from standing out from the other stimuli. The rationale behind was to add acoustic variability to stimuli and therefore to encourage the participant to abstract away from low-level acoustic signals. Note, however, that this decision is not critical with regard to data analysis, as only data from target stimuli were included.

4.2.2. English stimuli

The target stimuli for the English version of the perception experiment were identical to those for the Mandarin version. The filler stimuli, on the other hand, were changed to [mi˥], [mi˩˧] (similar to English *me*), [ni˥], and [ni˩˧] (similar to English *knee*). The reason why [i˥] and [i˩˧] were not used was to avoid the use of letter *E* as one of the response options; it was preferable that all four response options were lexical items.

4.3. Procedure

In presenting the experimental procedure, I first go through the configuration and layout of response options in each trial, and then described the task involved. At a high level, the task was a forced-choice identification task, where the participant clicked on one word out of a choice of four.

4.3.1. Mandarin trial configuration

Experimental trials consisted of two trial types: targets and fillers, depending on whether the audio stimulus being played were from the target or filler set. Both trial types had as response options four Mandarin monosyllabic words. For the targets, the four response words were *pi1* 披, *pi4* 屁, *bi1* 逼, and *bi4* 闭, which differed from one another in stop voicing and lexical tone. Note that these words were also included in the production stimuli. The four options were placed at the four corners of a 600 px × 600 px square, with each option having a response area of a 50 px × 50 px square, as illustrated in Supplementary Figure 2. Furthermore, the relative positions of the four options were constrained in such a way that two words distinguished only in the voicing of onset (e.g., *pi1* vs. *bi1*) were always next to each other, so there were only 16 (4 sides × 4 possible positionings/side) possible trial option configurations. The 16 trial configurations were counterbalanced across participants at the time of testing (i.e., the counterbalance was not taken into account when participants' data was selected for analyses), and the same configuration was used throughout the course of experiment. The decision to maintain the same configuration was to prevent the participant from doing visual search, which might introduce additional cognitive load.

For the fillers, the four options were *yi1* 衣, *yi4* 意, *mi1* 咪, and *mi4* 密, which similarly differed in both onset and lexical tone. However, their positioning was not constrained in any manner, as the data collected in filler trials were not analyzed. This resulted in 24 (= 4!) possible configurations, and each participant was randomly assigned a configuration, which remained the same throughout the entire experiment.

4.3.2. English trial configuration

The experimental trials for English similarly consisted of target trials and filler trials. However, unlike the Mandarin version, the two trial types differed from each other only in the audio stimulus being played; that is, the same response layout was used for both trial types. This being the case came from the fact that English lacks lexical tone, so it was impossible to have a response layout parallel to that in the Mandarin version. The trial configuration always had as response options four English words: *pea*, *bee*, *me*, and *knee*. The four words were arranged such that *pea* and *bee* were always only one edge away from each other (and as a consequence *me* and *knee* were likewise always next to each other)—the same constraint that phonological competitors in terms of stop voicing were always adjacent to each other. This resulted in 16 possible option configurations (4 sides × 4 arrangements/side), two of which are shown in Supplementary Figure 3. These 16 configurations were counterbalanced across participants at the time of testing, and the configuration remained unaltered within an experiment session.

4.3.3. Task procedure

The experiment procedure was the same for both the Mandarin and English versions of the experiment. The whole experiment took place online and was programmed in jsPsych (de Leeuw, 2015). Participants were encouraged to use a physical mouse and to wear headphones for the experiment, though they could also do the experiment with a touchpad and/or the built-in loud speakers on their computer. The experiment started with a short hearing test, where the participant had to select the quietest tone out of three tones differing in loudness. This test was challenging to do when *not* wearing headphones. They had to respond correctly in at least five out of six trials to pass the test.

The basic procedure followed that of Experiment 1 from Dale et al. (2007). During each trial, the four options were first presented for 500 ms to remind the participant of the word at each corner. Next, a black dot, the radius of which was 5 px, appeared in the center of the screen, which the participant had to click for the audio stimulus to be immediately presented. The function of this center dot was to ensure that the mouse cursor was reset to (approximately) the center. The participant then had a 3-s period to indicate their response by clicking one of the words.

Participants had to go through three blocks, with each block having the same tokens and differing only in the order in which the tokens were presented. To have a target-to-filler ratio of about 4:1, each block contained one repetition of target stimuli and seven repetitions of filler stimuli, resulting in a total of 126 (= 98 × 1 + 7 × 4) trials in each block. Three blocks were used to achieve a compromise between having as many trials as possible and limiting the duration of the experiment under 30 min. Between blocks the participant could take a self-timed break.

4.4. Additional participant inclusion criteria

As mentioned in Section 3.1, participants' performances in the perception experiment formed a part of the inclusion criteria. The purpose is to only include participants who actually paid attention during the experiment. This criterion was operationalized by first calculating by-participant "correct" percentage of responses for each language version, separated for target and filler trials. For the target trials in the Mandarin perception experiment, a correct trial was a target trial where the participant selected as the response a word whose tone matched the tonal contour of the audio stimulus. For the filler trials in the Mandarin experiment, a correct trial was a filler trial whose selected response word corresponded exactly to the audio stimulus (e.g., selecting *yi1* for [i˥]). For the target trials in the English version of the experiment, a correct trial was a target trial whose response was either *pea* or *bee*. For the filler trials in the English experiment, a correct trial was defined as a filler trial which had *me* or *knee* as the response, taking into account the fact that the bilabial and alveolar nasal onsets in the filler stimuli were perceptually confusable. For a participant who completed both English and Mandarin perception experiments, four percentage scores were computed—% correct for targets in Mandarin perception, % correct for fillers in Mandarin perception, % correct for targets in English perception, and % correct for fillers in English perception. For each participant, an average correct percentage across the four language/trial type combinations was computed. Participants were then ranked based on the average correct percentage in a descending order, and the data from the top 25 participants was included in the analyses. A *post-hoc* analysis shows that these included participants had an average correct percentage of at least 90%.

4.5. Omitted data

For both Mandarin and English versions of the perception experiment, only the response data from the target trials were considered. Additionally, only the "correct" target trials, as defined in Section 4.4 above, were included in the analyses. Altogether, 216 (129 Tone 1 tokens and 87 Tone 4 tokens) out of

7,350 target trials were removed from the Mandarin experiment, and 59 (29 Tone 1 tokens and 21 Tone 4 tokens) out of 7,350 target trials were removed from the English experiment.

4.6. Statistical analyses

A variant of logistic regression was used to derive the perceptual weight for post-stop F0. In all the models, participants' responses were modeled as a function of VOT, post-stop F0, and tonal categories. The coefficient of the post-stop F0 variable was then used as its perceptual weight. Similar to the production models, all models were fitted with Bayesian mixed-effects models using `CmdStanR` (Gabry and Češnovar, 2021).

4.6.1. Variables

Before being fed into the analyses, the two continuous predictor variables—**VOT** and **post-stop F0**—were z-transformed with respect to the original sequence (e.g., the VOT value of 0 was consistently mapped to [0 − mean(0, 13, 27, 40, 53, 67, 80)]/sd(0, 13, 27, 40, 53, 67, and 80) = −1.39, regardless of listener). The variable **tone** was sum-coded with TONE 1 and TONE 4 being coded with 1 and −1, respectively. The default level for the response was always unaspirated (i.e., the unaspirated response was coded with 0, and the aspirated response was coded with 1), so a positive coefficient for a given predictor variable means that higher values of this dimension elicit more voiceless responses in listeners than lower values.

4.6.2. Model structure

Listeners' responses were assumed to be generated by a mixture of two different sources: one source was the logistic function of terms formed with the predictors, and the other was sheer randomness or guessing due to the listener not paying attention or accidentally making a mistake, that is, the response came from one of the four options being selected by chance (Kruschke, 2015). Formally, each response had a chance, γ, of being generated by the guessing process, and, with probability $1-\gamma$, the response came from the logistic function of the predictor:

$$\text{aspirated response} \sim \text{bernoulli}\left(\gamma \cdot \frac{1}{4} + (1-\gamma)\cdot \text{logistic}\left(\beta_0 + \sum_i \beta_i x_i\right)\right).$$

Model fitting thus involved estimating the guessing probability γ along with the logistic parameters, β_i, which were taken to represent the weight given to each dimension in categorization. Bayesian hierarchical models were employed to derive a posterior probability distribution for each parameter. The full model consisted of two submodels with the same parameterization and predictors: one submodel predicted listeners' responses in the Mandarin mode while the other submodel predicted listeners' responses in the English mode, and the two submodels were tied together by correlating all logistic parameters with one another in a multinormal distribution. A guessing probability was estimated for each listener in each language mode independently. Logistic parameters were parameterized such that each was decomposed into a fixed-effect part, corresponding to the weight at the population level, and a random-effect part, representing the adjustment for each listener.

Each model used 4,000 samples across four Markov chains and was fit with a regularizing prior of Normal($\mu = 0$, $\sigma = 10$) for the fixed-effect estimates. An Exponential($r = 1$) distribution was used as the prior for listener-specific adjustments. Correlations among listener-specific adjustments used the LKJ prior with $\xi = 1$. The guessing probability for each listener in each language had a uniform prior between 0 and 1. All models showed no divergent transitions, and sampling chains were well-mixed (i.e., all $\hat{R} < 1.01$). The detailed mathematical specifications for the final model can be found in the Supplementary material.

4.6.3. Candidate models

Similar to the statistical models for production data, candidate models for perceptual performance reflected both prior knowledge and a compromise between complexity and predictive accuracy. Given that VOT is the primary cue for the stop voicing contrast in Mandarin and English, all the models in the comparison had VOT automatically included, with the simplest model containing VOT as the sole predictor. Built off this simplest models were candidates with increasing complexity introduced by terms involving post-stop F0 and tone. The full list of models considered is listed in Table 4.

4.7. Results: Perceptual weights of post-stop F0

The response patterns across different VOTs, post-stop F0s, tones, and experiment versions are shown in Figure 6. The ELPD-LOO mean and standard error for each candidate model are listed in Table 4, and the model comparison results among the candidate models are detailed in Supplementary Table 5.

Model comparison indicated the importance of post-stop F0 and tone in predicting listeners' categorization performances (M1 vs. M2 and M3 vs. M4 for post-stop F0; M1 vs. M3 and M2 vs. M4 for tone). However, including interaction terms between any pairs of the cues did not lead to substantial increase in

TABLE 4 Candidate perceptual models considered in model comparison, with their ELPD-LOO means and standard errors.

Model	ELPD-LOO mean	ELPD-LOO standard error	Predictors
M1	−1419.5	52.3	VOT
M2	−1366.2	51.3	VOT + F0
M3	−1395.4	52.0	VOT + tone
M4 (final)	−1340.5	51.4	VOT + F0 + tone
M5	−1334.6	51.5	VOT + F0 + tone + F0 × VOT
M6	−1325.4	51.7	VOT + F0 + tone + F0 × tone
M7	−1326.0	51.8	VOT + F0 + tone + F0 × VOT + F0 × tone
M8	−1327.9	52.1	VOT + F0 + tone + F0 × VOT + F0 × tone + VOT × tone

predictive accuracy. For this reason, M4 was selected as the final model, and subsequent discussion was made on the basis of M4.

4.7.1. Population results

The marginal posterior distributions for population-level effects from M4 are summarized in Table 5. All predictors, including the intercepts, had an effect on categorization. The cue of most interest here is post-stop F0, but for completeness, the results for other dimensions are also briefly discussed. On the basis of the fact that the 89% CrIs for post-stop F0 did not contain 0 in both Mandarin and English (Mandarin: 89% CrI = [0.30, 0.75]; English: 89% CrI = [0.64, 1.14]), post-stop F0 was judged to be a cue for stop voicing in both languages. However, the weight assigned to this cue was language-dependent, as evidenced by the 89% CrI of difference in post-stop F0 weights occupying only negative values (89% CrI = [−0.67, −0.04]). In particular, listeners relied on post-stop F0 more when the stimuli were presented as English words than when the exactly same stimuli were perceived as Mandarin words. The magnitude of the intercept was indicative of the location of category boundary: a positive intercept meant there were more aspirated responses in general, which translated to an early boundary within the range of values considered. This can be clearly seen in Figure 6, where the category boundary in terms of VOT (i.e., the VOT value where the proportion of aspirated responses is 0.5) occurs before the midpoint of the VOT continuum. Also, the intercept seemed stable across participants' Mandarin and English categorization performances. VOT, as expected, was the strongest cue for the voicing decision, and its weight was comparable across languages. Finally, Tone 1 stimuli seemed to trigger more aspirated responses to a similar degree in both languages.

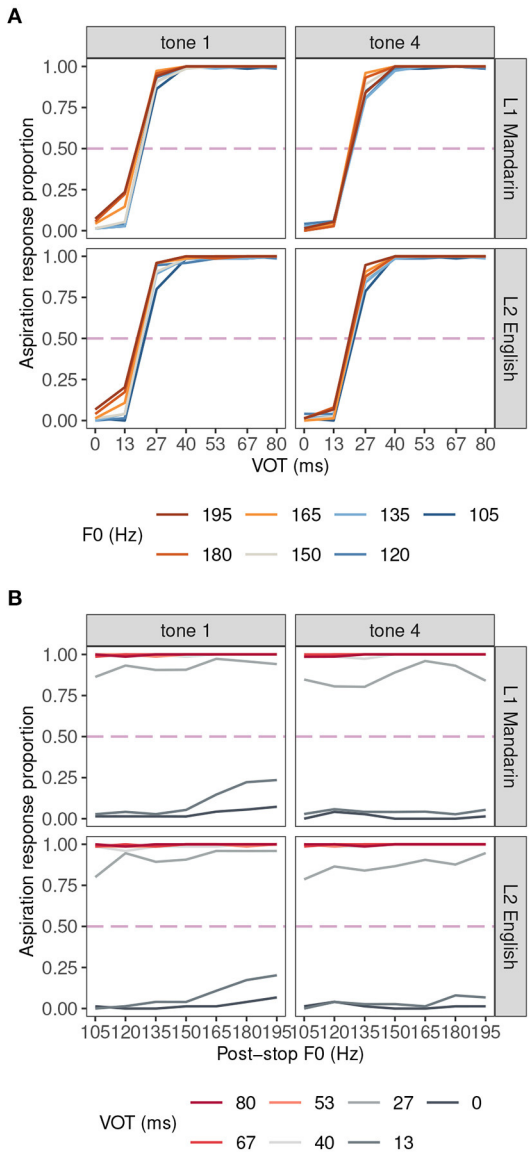

FIGURE 6
Line charts of Mandarin-English bilinguals' aggregated categorization of word-initial stops in each language, shown as a function of VOT, post-stop F0, and tone. **(A)** With VOT on the *x*-axis. **(B)** With post-stop F0 on the *x*-axis to highlight its effect on categorization.

4.7.2. Individual results

The guessing probability estimated for each listener in each language is plotted in the Supplementary Figure 4. Overall, the guessing probabilities were very low, with 24 out of 25 listeners having a mean guessing probability below 5% in either language and only one listener (i.e., participant 12) having a value of around 10% for the English task.

Individual listeners' weights for various cues, which are equal to the coefficient estimates of the corresponding acoustic dimensions, and the weight differences in these cues across

TABLE 5 Marginal posterior summary for key population-level parameters from M4.

Parameter	Mean	SD	89% CrI	p(dir.)
$\text{intercept}_{\text{Man}}$	9.14	0.68	[8.11, 10.28]	$p(\beta > 0) = 1.00$
VOT_{Man}	13.81	1.07	[12.19, 15.63]	$p(\beta > 0) = 1.00$
F0_{Man}	0.53	0.14	[0.30, 0.75]	$p(\beta > 0) = 1.00$
tone_{Man}	0.54	0.12	[0.35, 0.73]	$p(\beta > 0) = 1.00$
$\text{intercept}_{\text{Eng}}$	9.88	0.73	[8.81, 11.07]	$p(\beta > 0) = 1.00$
VOT_{Eng}	15.08	1.11	[13.42, 16.90]	$p(\beta > 0) = 1.00$
F0_{Eng}	0.89	0.16	[0.64, 1.14]	$p(\beta > 0) = 1.00$
tone_{Eng}	0.42	0.12	[0.22, 0.61]	$p(\beta > 0) = 1.00$
$\text{intercept}_{\text{Man}} - \text{intercept}_{\text{Eng}}$	−0.74	0.92	[−2.16, 0.74]	$p(\beta < 0) = 0.78$
$\text{VOT}_{\text{Man}} - \text{VOT}_{\text{Eng}}$	−1.28	1.41	[−3.43, 1.08]	$p(\beta < 0) = 0.81$
$\text{F0}_{\text{Man}} - \text{F0}_{\text{Eng}}$	−0.36	0.20	[−0.07, −0.04]	$p(\beta < 0) = 0.97$
$\text{tone}_{\text{Man}} - \text{tone}_{\text{Eng}}$	0.12	0.17	[−0.15, 0.39]	$p(\beta > 0) = 0.75$
$\rho_{\text{intercept}_{\text{Man}},\text{intercept}_{\text{Eng}}}$	0.41	0.21	[0.04, 0.74]	$p(\beta > 0) = 0.96$
$\rho_{\text{VOT}_{\text{Man}},\text{VOT}_{\text{Eng}}}$	0.52	0.19	[0.20, 0.79]	$p(\beta > 0) = 0.99$
$\rho_{\text{F0}_{\text{Man}},\text{F0}_{\text{Eng}}}$	0.34	0.28	[−0.15, 0.73]	$p(\beta > 0) = 0.88$
$\rho_{\text{tone}_{\text{Man}},\text{tone}_{\text{Eng}}}$	0.10	0.33	[−0.44, 0.62]	$p(\beta > 0) = 0.62$
$\rho_{\text{VOT}_{\text{Man}},\text{F0}_{\text{Man}}}$	−0.33	0.24	[−0.70, 0.08]	$p(\beta < 0) = 0.90$
$\rho_{\text{VOT}_{\text{Eng}},\text{F0}_{\text{Eng}}}$	−0.06	0.27	[−0.50, 0.38]	$p(\beta < 0) = 0.59$
$\rho_{\text{tone}_{\text{Man}},\text{F0}_{\text{Man}}}$	0.20	0.31	[−0.32, 0.66]	$p(\beta > 0) = 0.75$
$\rho_{\text{tone}_{\text{Eng}},\text{F0}_{\text{Eng}}}$	0.11	0.30	[−0.40, 0.59]	$p(\beta > 0) = 0.65$
$\rho_{\text{tone}_{\text{Man}}-\text{tone}_{\text{Eng}},\text{F0}_{\text{Man}}-\text{F0}_{\text{Eng}}}$	−0.11	0.34	[−0.67, 0.42]	$p(\beta < 0) = 0.63$

languages are visualized in Figure 7. Again, the results regarding the cue weight for post-stop F0 are discussed first, as it is the dimension of interest here; the results for other cues are also summarized in passing for completeness.

As shown in the [post-stop F0] panel of Figure 7A, though the 89% CrI for the post-stop F0 weight did cross 0 for some listeners, all listeners had a positive mean weight for the post-stop F0 cue for both languages, signifying that, generally speaking, the chance the aspirated response was selected went up with an increasing post-stop F0. Comparing the weights of this cue across languages (Figure 7B), all but one listener (i.e., participant 25) had a higher mean weight in English than in Mandarin; however, because of the relatively large uncertainty surrounding the estimated weight values, the 89% CrI for the *difference* between the weights still contained 0 for all participants. In spite of this "non-significant" result, the trend seemed robust and echoed the population-level pattern in terms of the direction of the effect. Another way to understand the cue is to examine whether the cue use is consistent across languages at the individual level by correlating the weights from the two language contexts. In fact, the correlation information can be directly read off from the fitted model and is summarized in the last few row in Table 5 and visualized in Figure 8. As can be seen in Figure 8C, there was a weak positive correlation of this cue across languages ($\bar{\rho} = 0.34$, 89% CrI = [−0.15, 0.73], $p(\rho > 0) = 0.88$), though the 89% CrI for this correlation also spilled to the negative side, probably due to the small number of participants, which was not effective in constraining the uncertainty when the correlation was weak.

For the intercepts, which were connected with the location of category boundary, even though individual listeners varied with respect to the boundary location, the location was relatively stable within a listener, as evidenced from Figure 8A and from the positive 89% CrI of the correlation coefficient ($\bar{\rho} = 0.41$, 89% CrI = [0.04, 0.74], $p(\rho > 0) = 0.96$). The same story could be stated for the VOT cue: individuals varied in a structured way, with the cue use being stable within the same individual across contexts ($\bar{\rho} = 0.52$, 89% CrI = [0.20, 0.79], $p(\rho > 0) = 0.99$). As for tone, it seemed that, for most listeners (19 out of 25), the effect of Tone 1 stimuli eliciting more voiceless responses was stronger in Mandarin than in English, though the difference was not particularly big.

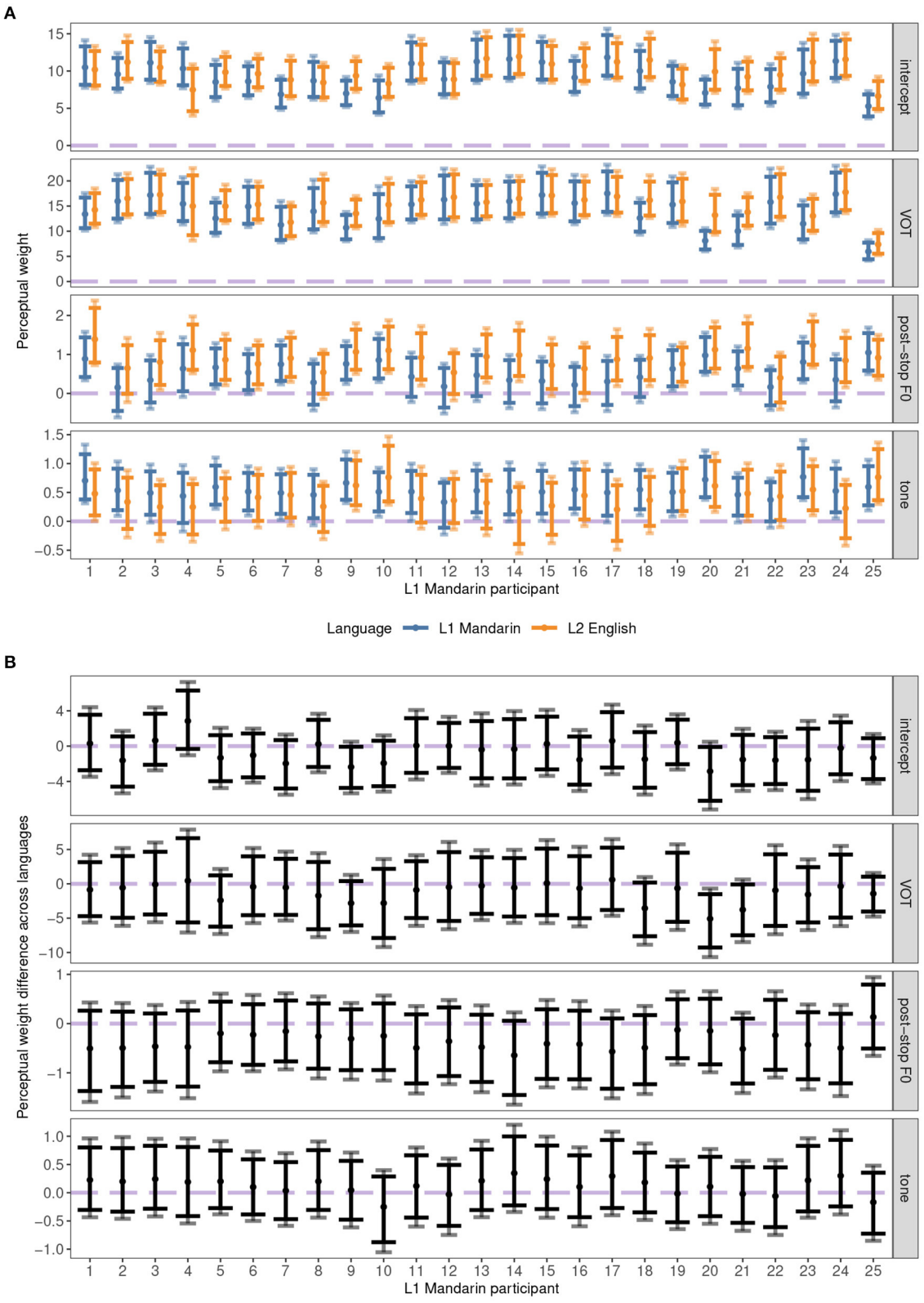

FIGURE 7
Individuals' estimated weights from the perceptual model. **(A)** Distributions of individual weights along various dimensions for Mandarin and English. **(B)** Differences in cue weights along the same dimension across languages. In both figures, posterior means are represented by the dots. The 89% CrIs are marked by the inner error bars, while the 95% CrIs are marked by the outer error bars.

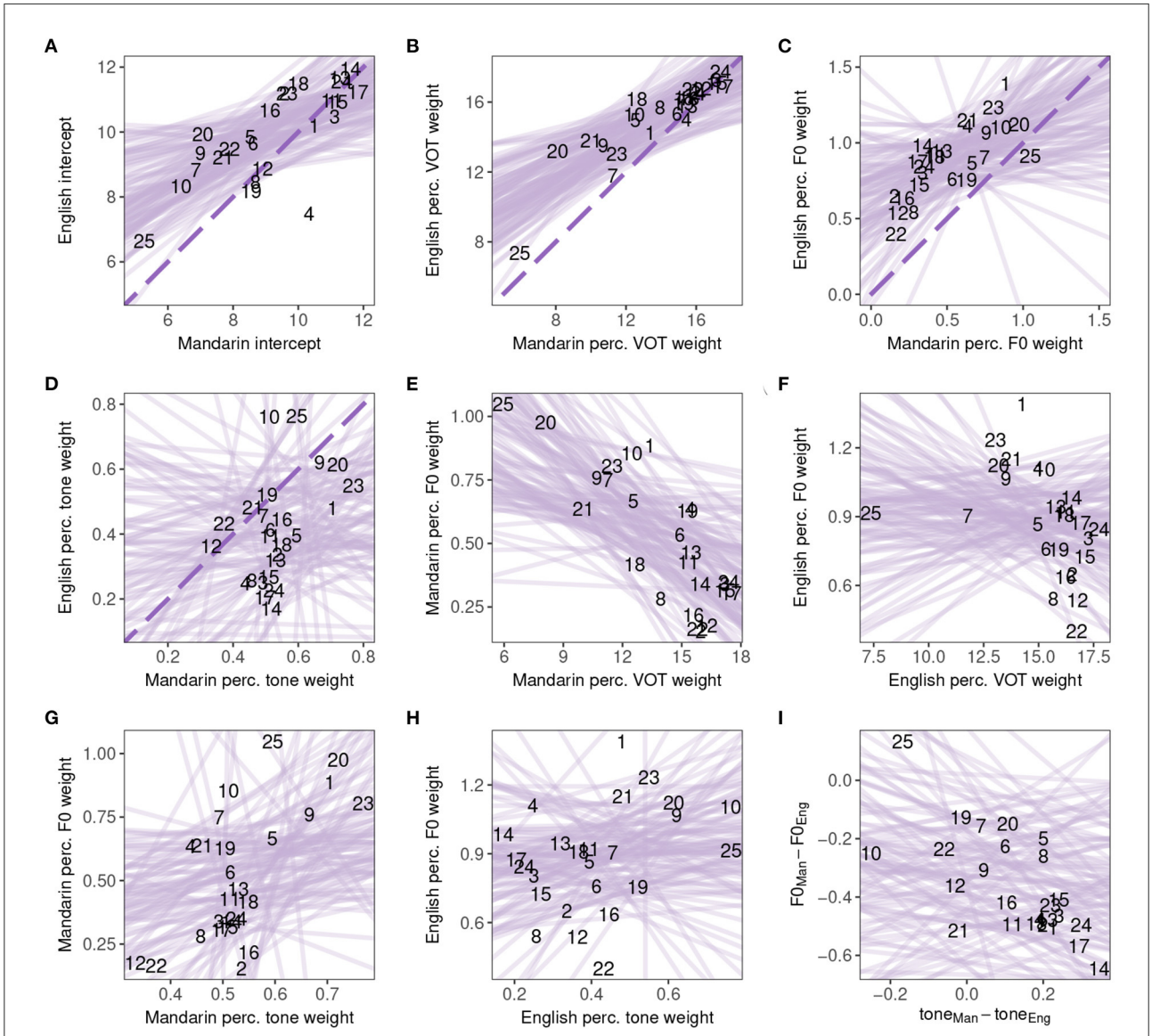

FIGURE 8
Scatter plots showing relationships (or lack thereof) between various cues. **(A)** Intercepts, which are related to category boundaries, across languages. **(B)** VOT weights across languages. **(C)** Post-stop F0 weights across languages. **(D)** Tone weights across languages. **(E)** F0 vs. VOT in Mandarin. **(F)** F0 vs. VOT in English. **(G)** F0 vs. tone in Mandarin. **(H)** F0 vs. tone in English. **(I)** Differences in post-stop F0 weights vs. differences in tone weights. Solid lines represent 100 regression lines fitted with 100 posterior draws, to show the direction and uncertainty of the correlation. The dashed line in **(A–D)** is $y = x$, where the intercept or VOT / post-stop F0 / tone weight for Mandarin equals that for English.

4.8. Comparing individual post-stop F0 weights across production and perception

Given that population-level correspondences between production and perception alone cannot be taken as evidence for a causal link—if there is a (direct or indirect) causal link between the modalities, it should surface on an individual level (Schertz et al., 2020). It is therefore expected that the weight of a given acoustic dimension on a speaker's production would predict the weight assigned to that dimension in the same speaker's perception. To test this hypothesis empirically, two models, separated for each language but otherwise sharing the same structure, were fit using both production and perception data. Each model had two submodels: one estimated individual production weights based on Cohen's *d*, and the other estimated individual perceptual weights based on the beta-coefficient for F0 in the logistic regression model. The two submodels were tied together by a common covariance matrix used to model individual-level variances. The

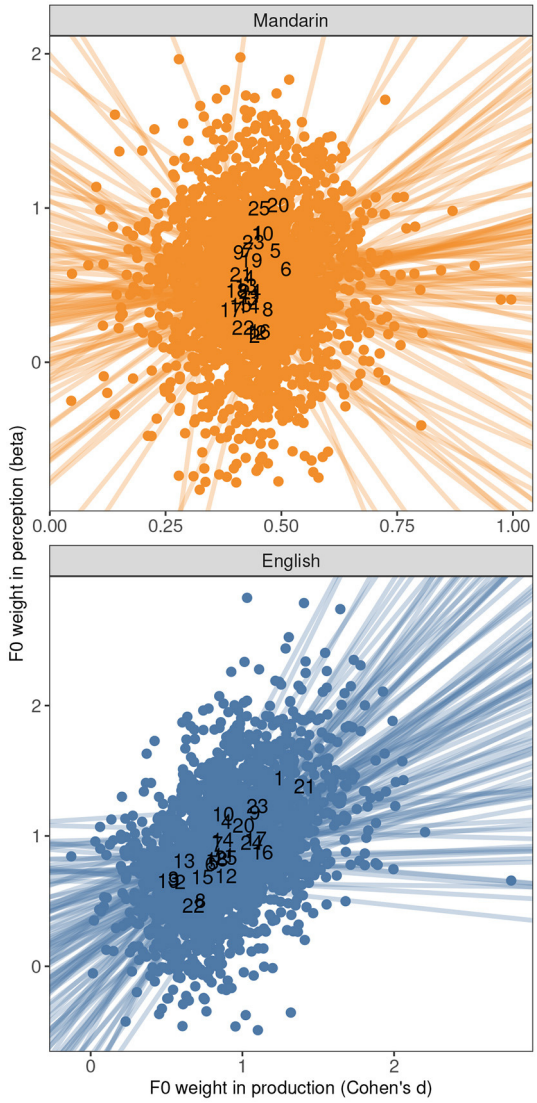

FIGURE 9
Post-stop F0 weights across perception and production in Mandarin and English. The numbers represents the posterior means for individual participants. Each dot is a posterior draw for a participant. The solid lines represent 100 regression lines fitted with 100 posterior draws, to show the direction and uncertainty of the correlation.

mathematical specification for the models can be found in the Supplementary material. Figure 9 shows individual perceptual weights plotted against the corresponding production weights. The results of correlation analyses were dependent on the language, with little evidence of correlation across modalities for Mandarin ($\bar{\rho}$ = 0.49, 89% CrI = [−2.93, 4.18], $p(\rho > 0)$ = 0.61) but weak evidence for a positive correlation for English ($\bar{\rho}$ = 0.72, 89% CrI = [−0.15, 1.51], $p(\rho > 0) = 0.93$).

5. Discussion

5.1. Summary of results

The current study explores the ambiguity of F0 in Mandarin through L1 Mandarin-L2 English bilinguals' production and perception of the stop voicing contrast in their L1 and L2. The results from the conducted experiments are summarized in Table 6, which ties them back to the hypotheses and predicted results listed in Table 1, and discussed below. At the population level, these results largely echoed a recent work by Guo (2020).

In both their Mandarin and English productions, the post-stop F0 following an aspirated stop tended to be higher than that following an unaspirated stop, and unaspirated stops in turn induced a higher F0 than sonorants. In addition, the extent to which post-stop F0 was differentiated between the aspirated and unaspirated categories hinged on the language and lexical tone: comparing English with Mandarin (which was represented as an average between Tone 1 and Tone 4 in this study), English supported a bigger post-stop F0 difference; contrasting Tone 1 and Tone 4 in Mandarin, Tone 4, which was realized with a higher F0 register phonetically, also sustained a slightly greater post-stop F0 distinction. The production weights for post-stop F0 across languages was also reflective of the finding above: post-stop F0 assumed a larger weight in English than in Mandarin, both at the population level and for most individuals (19 out of 25 speakers). These findings therefore support the view that post-stop F0 perturbation is not necessarily intrinsic to the articulatory system.

In perception, post-stop F0 was also used as a cue for stop voicing by the same L1 Mandarin-L2 English participants when put in either a Mandarin or an English context. However, the language context modulated the weight such that post-stop F0 carried more weight when the stimuli were presented as English words than when the same stimuli were presented as Mandarin words. This language-conditioned change in cue weighting was statistically well-supported at the population level, but, at the individual level, because of fewer data points (i.e., the same stimuli were only repeated three times for each participant), the model was less confident. Nonetheless, almost all individuals (24 out of 25) followed the population trend as far as posterior means were concerned. Overall, the patterns revealed in the perception experiment are supportive of the claim that L2 learners can adjust the use of a cue in different language contexts.

Compared across production and perception, on a population level, a higher production weight for post-stop F0 mapped to a higher perceptual weight for the same cue. This is reflected in the bilinguals' relying more on post-stop F0 to contrast stop voicing in English than in Mandarin across modalities. On an individual level, on the other hand, an individual's production weight did not reliably predict the same individual's perceptual weight,

TABLE 6 Predicted and actual production and perception results under difference hypotheses.

Production		
Hypotheses	**Predicted production results**	**Match actual results?**
Post-stop F0 purely due to physiological / aerodynamic reasons (e.g., Ladefoged, 1967; Ohala and Ohala, 1972; Kohler, 1984) or total transfer of post-stop F0 cue use in Mandarin to English, as prediced by the SLM and PAM-L2	Post-stop F0 difference the same in Mandarin and English tokens	No
Post-stop F0 partially subject to active controlling (Kingston and Diehl, 1994)	The extent of post-stop F0 difference might depend on the language (i.e., larger in English than in Mandarin)	Yes. Post-stop F0 difference between aspirated and unaspirated stops was bigger in English than in Mandarin at the population level and for 19 (out of 25) speakers.
Perception		
Hypotheses	**Predicted perception results**	**Match actual results?**
Transfer of the Mandarin cue-weighting strategy to English, as predicted by the SLM and PAM-L2	Post-stop F0 weights the same across Mandarin and English	No.
Flexibility in cue use: attributing variation in post-stop F0 partially to lexical tone and partially to stop voicing in Mandarin, but only to stop voicing in English	Post-stop F0 weights depend on the language context (i.e., a higher weight in English than in Mandarin)	Yes. Post-stop F0 carried more weight in English than in Mandarin at the population level. The model was less confident at the individual level, though the trend was the same as the population result for 24 out of 25 listeners.

at least for post-stop F0 with the adopted metrics. This mismatch therefore suggests at least some independence of the two modalities.

5.2. Flexibility of cue-weighting across L1 and L2

The findings from the experiments show that bilinguals, even non-early/ non-simultaneous/non-child bilinguals, are able to dynamically adjust their cue-weighting strategies in facing different language contexts in production as well as perception. Prior demonstrations on bilinguals' ability to fine-tune the use of various acoustic dimensions concerned mainly simultaneous or early bilinguals (e.g., Antoniou et al., 2010, 2012; Gonzales and Lotto, 2013; Gonzales et al., 2019). However, as reviewed in Section 2.5, more recent works have suggested that late L2 learners are also capable of such a deed. The results from this study are in line with thse recent works in that Mandarin-English bilinguals shift the post-stop F0 weight in response to the current language mode. Crucially, however, this study also demonstrates bilinguals' capability to modulate the use of a secondary cue, as opposite to just the primary cue as in previous works.

5.3. Role of tone in post-stop F0

The fact that, in production, greater post-stop F0 difference was found in Tone 4, which was realized with a higher initial pitch than Tone 1, and that, in perception, Tone 1 syllables induced more aspirated responses, points to a potential role of tone identity in conditioning post-stop F0. In fact, previous works have documented such cases in production at least. For example, as mentioned in Section 2.2.2, Guo (2020) reports that F0 following an aspirated stop is higher only in Tone 1 and Tone 4 syllables (both of which begin with a high pitch register) while F0 following an *un*aspirated is higher in Tone 2 and Tone 3 syllables (both having a low initial register). Kirby (2018) investigates the post-stop F0 effects in two other tonal languages—Thai and Vietnamese—and finds that the greatest post-stop F0 effects for Thai are present in the high-falling tone environment, though the results from Vietnamese are less clear-cut. Even in non-tonal languages,

post-stop F0 difference is most prominent in high-pitch, focused conditions (Hanson, 2009; Kirby and Ladd, 2016). The enlargement of post-stop F0 difference in high-pitch contexts across tonal and non-tonal languages suggests that a general, language-independent explanation in terms of F0 control might be responsible, and more research is needed to elucidate this hypothesis.

With respect to perception, a careful inspection of Figure 6 reveals that increased aspirated responses in Tone 1 tokens resulted largely from higher post-stop F0 values in Tone 1 provoking more aspirated responses when VOT was ambiguous (i.e., when VOT was around 13 ms). A possible explanation for why Tone 1, as compared with Tone 4, led to such an effect is that it is not just the initial value of F0 that matters; the listener also tracks changes in F0 slope throughout the syllable, and such changes also contribute to the perception of F0. In the context of the current perception experiment, all Tone 1 tokens end with a tailing flat F0 contour, which might enhance the percept of the initial drop in F0, whereas the falling F0 contour in Tone 4 tokens might perceptually offset the initial drop in F0, resulting in the change in F0 being less noticeable. Another explanation is that since Tone 4 syllables tend to have a higher initial F0 in production than Tone 1 syllables, Mandarin listeners might require an acoustically higher initial F0 value in Tone 4 tokens to judge a token as starting with a high F0. Of course these speculations await more investigation.

Related to changes in F0 slope is the question, as pointed out by a reviewer, of whether the observed effect of post-stop F0 is induced by vowel-onset F0 height or by the F0 contour within the range of manipulation (i.e., from vowel onset to the 35% mark of the vowel). As can be seen in Figure 5B, the manipulation of F0 in this study conflates vowel-onset F0 height and F0 contour. For instance, for F0 manipulation in both Tone 1 and Tone 4 tokens, a higher vowel-onset F0 is associated with a more positive F0 contour. As reviewed in Section 2.2.3, both F0 height and F0 contour contribute to perception of various pitch events. It is therefore possible that both vowel-onset F0 and F0 contour drive the perception of an aspirated stop for a high post-stop F0. One possible future direction is to tease apart the respective influence of the two manipulations.

5.4. A trade-off between post-stop F0 and tone?

The fact that the post-stop F0 weight is diminished in the Mandarin context across both production and perception raises the question of whether the lost weight in post-stop F0 is transferred to other dimensions, with the most obvious candidate being tonal category. In what follows, I discuss the case with production first before moving on to perception.

The question about the existence of a trade-off between post-stop F0 and tone is tied to the debate of whether tone attenuates the degree of post-stop F0 difference. As mentioned in Section 2.2.2, whereas there are some studies that point to a positive direction (e.g., Gandour, 1974; Hombert, 1978), large magnitudes of post-stop F0 difference have also been observed in tonal languages (e.g., Phuong, 1981; Shimizu, 1994; Xu and Xu, 2003; Francis et al., 2006). In the current study, the Mandarin-English bilinguals' respective language productions do conform to the former pattern at the population level. However, not every speaker matches the population-level trend, with some speakers producing the post-stop F0 effect to a similar degree in both languages. The results presented here thus agree with Kirby's (2018) observation that attenuation of post-stop F0 effect in tone languages depends on speaker-specific implementation of laryngeal maneuvers to distinguish voicing and tone.

With respect to perception, if, as described in Section 2.6, it is indeed the case that, in interpreting the audio stimuli as Mandarin words, Mandarin-English bilinguals attribute the variation in post-stop F0 partially to the lexical tones in the language, and that in treating the stimuli as English words, they ascribe the variation to stop voicing, then it is expected the loss in post-stop F0 weight from Mandarin to English to be accompanied by an increase in tone weight. Looking at Table 5, which shows the results at the population level, it seems the loss in post-stop F0 is indeed accompanied by an increase in tone weight, though the model is not as confident in the increase in tone weight as in the decrease in post-stop F0 weight. At the individual level, the panels for post-stop F0 and tone in Figure 8 also appear to suggest that for many participants, a drop in post-stop F0 weight is compensated by a rise in tone weight, and that those who have a bigger drop tend to have a sharper rise as well (notice the apparent negative correlation in Figure 8I when the changes along these two dimensions are plotted against each other), at least as far as the posterior mean is concerned. However, the correlation coefficient estimated from posterior samples does not back up this hypothesis (as shown by the lines going into different directions in Figure 8I). Therefore, it is still inconclusive as to whether there is a trade-off relation between post-stop F0 and tone in perception.

5.5. Production-perception interface

As shown in Section 4.8, even though the use of post-stop F0 is mirrored across production and perception at the population level, the link between the two modalities at the individual level seems to be less robust. While there is weak evidence for a positive correlation between the production and perceptual weights for English, such a correlation is missing for Mandarin. This observation raises the question as to the cause of this asymmetry. One possible answer might be that post-stop F0

is an unreliable cue for phonological voicing in Mandarin. For instance, looking at Figure 7A, almost all individuals use post-stop F0 to a lesser degree as a cue for voicing in Mandarin than in English; for many, the model indicates only very weak evidence for the use of post-stop F0 as a cue. This lack of robustness in the perpetual use of post-stop F0 in Mandarin can be understood in the context of production results from previous studies. Recall from the review in Section 2.2.2 that conflicting findings have been reported regarding the direction of post-stop F0 perturbation in Mandarin. These findings might be suggestive of an inconsistent patterning between post-stop F0 and voicing in Mandarin, and/or large individual variation in this patterning due to dialects, L2 influences, etc. The net result is that Mandarin listeners learn to downweight the post-stop F0 cue as it is only marginally useful in signaling voicing. In other words, the lack of link between production and perception in Mandarin at the individual level comes about because listeners downweight the use of post-stop F0 in the face of potentially conflicting cue use in ambient speech, even though they might produce post-stop F0 in a consistent manner. This explanation is therefore in line with the proposal put forth by Beddor (2015) and Samuel and Larraza (2015) that individuals command a more flexible perceptual proficiency than their production repertoire in order to accommodate potentially large between-speaker variation.

It is worth pointing out that, among the studies that sought to establish individual-level correlation in cue use, a lack of relationship seems to be the norm. For instance, null results have been reported for VOT and F0 in English (Shultz et al., 2012), VOT, F0, closure duration, and F1 onset for English and Spanish (Schertz, 2014), or VOT, F0, and closure duration in L1 Korean and L2 English (Schertz et al., 2015), among other studies that used fairly standard paradigms similar to the one employed in this study. These studies also have in common estimating correlations from individuals' empirical mean cue weights. Such approaches disregard uncertainty surrounding the estimates, so the apparent correlation (or lack of correlation) might not be reliable. To properly account for the uncertainty requires fitting both production and perception data with a single model, and the resulting correlation might not agree with the apparent correlation based on means (M. Sonderegger, personal communication, May 20, 2022). Future research will therefore benefit from directly modeling the uncertainty.

6. Conclusion

The current work examines whether and how L1 Mandarin-L2 English bilinguals use post-stop F0 as a cue for stop voicing across production and perception in Mandarin as well as English contexts. The production results show that F0 is actively used to encode both tonal and voicing distinctions in their Mandarin tokens, and that voicing distinctions are likewise embedded with post-stop F0 in English tokens. In perception, the bilinguals are also able to extract voicing information from post-stop F0 (in the same direction as observed in production) in both languages, even when post-stop F0 is integrated in the overall pitch contour, which they need to monitor in order to identify the lexical tone. Crucially, the reliability of post-stop F0 in signaling the voicing contrast and the extent to which the bilinguals lean on post-stop F0 for voicing perceptually are language-specific, such that production and perceptual weights for post-stop F0 are greater in the English context. However, a positive correlation between production and perceptual weights at the individual level is only observed for English, but not for Mandarin. This lack of correlation in Mandarin is interpreted as reflecting Mandarin listeners' flexible perceptual strategies in response to large individual variability in the direction of post-stop F0 perturbation in Mandarin.

Data availability statement

The datasets and analysis script for this study can be found below: https://osf.io/kw8ph/.

Ethics statement

The studies involving human participants were reviewed and approved by UBC Behavioural Research Ethics Board. The participants provided their written informed consent to participate in this study.

Author contributions

The research and writing were done by RY-HL.

Acknowledgments

My many thanks go to Kathleen Currie Hall, Molly Babel, and Márton Sóskuthy for their support and feedback throughout this project.

Conflict of interest

The author declares 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.864127/full#supplementary-material

References

Abramson, A. S., and Lisker, L. (1985). "Relative power of cues: F0 shift versus voice timing," in *Phonetic linguistics: Essays in honor of Peter Ladefoged*, ed V. A. Fromkin (New York, NY: Academic Press), 25–33.

Allen, J. S., and Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics ofmonosyllabic words. *J. Acoust. Soc. Am*. 106, 2031–2039. doi: 10.1121/1.427949

Amengual, M. (2021). The acoustic realization of language-specific phonological categories despite dynamic cross-linguistic influence in bilingual and trilingual speech. *J. Acoust. Soc. Am*. 149, 1271–1284. doi: 10.1121/10.0003559

Antoniou, M., Best, C. T., Tyler, M. D., and Kroos, C. (2010). Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *J. Phon*. 38, 640–653. doi: 10.1016/j.wocn.2010.09.005

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. Phon*. 40, 582–594. doi: 10.1016/j.wocn.2012.05.005

Barnes, J., Veilleux, N., Brugos, A., and Shattuck-Hufnagel, S. (2010). "The effect of global F0 contour shape in the perception of tonal timing contrasts in American English intonation," in *Proceedings of Speech Prosody 2010* (Chicago, IL), 1–4.

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

Beddor, P. S. (2015). "The relation between language users' perception and production repertoires," in *Proceedings of the 18th International Congress of Phonetic Sciences* (Glasgow: University of Glasgow), 1–9.

Best, C. T., and Tyler, M. D. (2007). "Nonnative and second-language speech perception: commonalities and complementarities," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, eds O.-S. Bohn and M. J. Munro (Amsterdam: John Benjamins Publishing Company), 13–34.

Blicher, D. L., Diehl, R. L., and Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone2/Tone3 distinction: evidence of auditory enhancement. *J. Phon*. 18, 37–49. doi: 10.1016/S0095-4470(19)30357-2

Boersma, P., and Weenink, D. (2021). *Praat: Doing Phonetics by Computer [Computer Program], Version 6.1.38*. Retrieved from: https://www.fon.hum.uva.nl/praat/

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: a probabilistic programming language. *J. Stat. Softw*. 76, 1–32. doi: 10.18637/jss.v076.i01

Casillas, J. V., and Simonet, M. (2018). Perceptual categorization and bilingual language modes: assessing the *double phonemic boundary* in early and late bilinguals. *J. Phon*. 71, 51–64. doi: 10.1016/j.wocn.2018.07.002

Chen, Y. (2011). How does phonology guide phonetics in segment-*F0* interaction? *J. Phon*. 39, 612–625. doi: 10.1016/j.wocn.2011.04.001

Clayards, M. (2018). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica* 75, 1–23. doi: 10.1159/000448809

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cole, J., Kim, H., Choi, H., and Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: evidence from Radio News speech. *J. Phon*. 35, 180–209. doi: 10.1016/j.wocn.2006.03.004

Dale, R., Kehoe, C., and Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Mem. Cogn*. 35, 15–28. doi: 10.3758/BF03195938

de Leeuw, J. R. (2015). jsPsych: a javascript library for creating behavioral experiments in a web browser. *Behav. Res. Methods* 47, 1–12. doi: 10.3758/s13428-014-0458-y

Dilley, L. C., and Heffner, C. C. (2013). The role of f0 alignment in distinguishing intonation categories: evidence from American English. *J. Speech Sci*. 3, 3–67. doi: 10.20396/joss.v3i1.15039

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. Phon*. 49:77–95. doi: 10.1016/j.wocn.2014.12.005

Ewan, W. (1976). *Laryngeal behavior in speech* (Ph.D. thesis). University of California, Berkeley, Berkeley, CA.

Flege, J. E. (1995). "Second language speech learning theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research, Chapter 8*, ed W. Strange (Timonium, MD: York Press), 233–277.

Flege, J. E. (2007). "Language contact in bilingualism: Phonetic system interactions," in *Laboratory Phonology 9*, eds J. Cole and J. I. Hualde (Berlin: Mouton de Gruyter), 353–381.

Fogerty, D., and Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *J. Acoust. Soc. Am*. 131, 1490–1501. doi: 10.1121/1.3676696

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *J. Phon*. 14, 3–28. doi: 10.1016/S0095-4470(19)30607-2

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

Francis, A. L., Kaganovich, N., and Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *J. Acoust. Soc. Am*. 124, 1234–1251. doi: 10.1121/1.2945161

Fulop, S. A., and Scott, H. J. M. (2021). Consonant voicing in the Buckeye corpus. *J. Acoust. Soc. Am*. 149, 4190–4197. doi: 10.1121/10.0005199

Gabry, J., and Češnovar, R. (2021). *cmdstanr: R Interface to 'CmdStan'*. Available online at: https://mc-stan.org/cmdstanr; https://discourse.mc-stan.org

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

Gandour, J. (1983). Tone perception in Far Eastern languages. *J. Phon*. 11, 149–175. doi: 10.1016/S0095-4470(19)30813-7

Gandour, J. T. (1978). "The perception of tone," in *Tone: A Linguistic Survey, Chapter 2*, ed V. A. Fromkin (New York, NY: Academic Press), 41–76.

Gao, J., and Arai, T. (2018). "F0 perturbation in a "pitch-accent" language," in *Proceedings of the Sixth International Symposium on Tonal Aspects of Languages* (Berlin), 56–60.

Gonzales, K., Byers-Heinlein, K., and Lotto, A. J. (2019). How bilinguals perceive speech depends on which language they think they're hearing. *Cognition* 182, 318–330. doi: 10.1016/j.cognition.2018.08.021

Gonzales, K., and Lotto, A. J. (2013). A Bafri, un Pafri: Bilinguals' pseudoword identifications support language-specific phonetic systems. *Psychol. Sci*. 24, 2135–2142. doi: 10.1177/0956797613486485

Guo, Y. (2020). *Production and perception of laryngeal contrasts in Mandarin and English by Mandarin speakers* (Ph.D. thesis). George Mason University, Fairfax, VA.

Han, M. S., and Weitzman, R. S. (1970). Acoustic features of Korean /P, T, K/, /p, t, k/ and /p^h, t^h, k^h/. *Phonetica* 22, 112–128. doi: 10.1159/000259311

Hanson, H. M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. *J. Acoust. Soc. Am*. 1, 425–441. doi: 10.1121/1.3021306

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099–3111. doi: 10.1121/1.411872

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica*. 33, 353–367. doi: 10.1159/000259792

Holt, L. L., and Lotto, A. J. (2006). Cue weighting in auditory categorization: implications for first and second language acquisition. *J. Acoust. Soc. Am.* 119, 3059–3071. doi: 10.1121/1.2188377

Hombert, J.-M. (1978). "Consonant types, vowel quality, and tone," in *Tone: A Linguistic Survey, Chapter 3*, ed V.A. Fromkin (New York, NY: Academic Press), 77–111.

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

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

Howie, J. M. (1976). *Acoustical Studies of Mandarin Vowels and Tones*. Cambridge: Cambridge University Press.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87, B47-B57. doi: 10.1016/S0010-0277(02)00198-1

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

Jun, S.-A. (1996). Influence of microprosody on macroprosody: a case of phrase initial strengthening. *UCLA Working Pap. Phonet.* 92, 97–116.

Kataoka, R. (2011). *Phonetic and cognitive bases of sound change*. (Ph.D. thesis). University of California, Berkeley, Berkeley, CA.

Keating, P., and Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *J. Acoust. Soc. Am.* 132, 1050–1060. doi: 10.1121/1.4730893

Kingston, J., and Diehl, R. L. (1994). Phonetic knowledge. *Language* 70, 419–454. doi: 10.1353/lan.1994.0023

Kingston, J., Diehl, R. L., Kirk, C. J., and Castleman, W. A. (2008). On the internal perceptual structure of distinctive features: the [voice] contrast. *J. Phon.* 36, 28–54. doi: 10.1016/j.wocn.2007.02.001

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

Kirby, J. P., and Ladd, D. R. (2016). Effects of obstruent voicing on vowel *F0*: evidence from "true voicing" languages. *J. Acoust. Soc. Am.* 140, 2400–2411. doi: 10.1121/1.4962445

Kohler, K. J. (1982). F_0 in the production of lenis and fortis plosives. *Phonetica* 39, 199–218. doi: 10.1159/000261663

Kohler, K. J. (1984). Phonetic explanation in phonology: the feature fortis/lenis. *Phonetica* 41, 150–174. doi: 10.1159/000261721

Koster, J., and McElreath, R. (2017). Multinomial analysis of behavior: statistical methods. *Behav. Ecol. Sociobiol.* 71, 1–14. doi: 10.1007/s00265-017-2363-8

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan, 2nd Edn*. London: Academic Press.

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics*. Oxford: Oxford University Press.

Lea, W. A. (1973). "Segmental and suprasegmental influences on fundamental frequency contours," in *Consonant Types and Tone: Southern California Occasional Papers in linguistics no. 1*, ed L. M. Hyman (Los Angeles, CA: The Linguistics Program; University of Southern California), 16–70.

Lee, B., and Sidtis, D. V. L. (2017). The bilingual voice: vocal characteristics when speaking two languages across speech tasks. *Speech Lang. Hear.* 20, 174–185. doi: 10.1080/2050571X.2016.1273572

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

Lehnert-LeHouillier, H. (2007). "The influence of dynamic F0 on the perception of vowel duration: cross-linguistic evidence," in *Proceedings of the 16th International Congress of Phonetic Sciences* (Saarbrücken), 757–760.

Leung, K. K. W., and Wang, Y. (2020). Production-perception relationship of Mandarin tones as revealed by critical perceptual cues. *J. Acoust. Soc. Am.* 147, EL301-EL306.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *J. Multivar. Anal.* 100, 1989–2001. doi: 10.1016/j.jmva.2009.04.008

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychol. Monogr. Gen. Appl.* 68, 1–13. doi: 10.1037/h0093673

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Liberman, M. (2014). *Consonant Effects on F0 in Chinese [Blog Post]*. Retrieved from: https://languagelog.ldc.upenn.edu/nll/?p=12902

Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Lang. Speech*. 29, 3–11. doi: 10.1177/002383098602900102

Liu, C., and Rodriguez, A. (2012). Categorical perception of intonation contrast: effects of listeners' language background. *J. Acoust. Soc. Am.* 131, EL427-EL433. doi: 10.1121/1.4710836

Lotto, A. J., Sato, M., and Diehl, R. L. (2004). "Mapping the task for the second language learner: the case of Japanese acquisition of /r/ and /l/," in *From Sound to Sense: 50+ Years of Discoveries in Speech Communication* (Cambridge, MA), C-181-C-186.

Luo, Q. (2018). *Consonantal effects on F0 in tonal languages* (Ph.D. thesis). Michigan State University, East Lansing, MI.

Ma, J. K.-Y., Ciocca, V., and Whitehill, T. L. (2006). Effect of intonation on Cantonese lexical tones. *J. Acoust. Soc. Am.* 120, 3978–3987. doi: 10.1121/1.2363927

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course With Examples in R and Stan*. Boca Raton, FL: CRC Press.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A. M., Jenkins, J. J., and Fujimura, O. (1975). An effect of linguistic experience: the discrimination of [r] and [l] by native speakers of Japanese and English. *Percept. Psychophys*. 18, 331–340. doi: 10.3758/BF03211209

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

Ohala, J. J. (1978). "Production of tone," in *Tone: A Linguistic Survey, Chapter 1*, ed V. A. Fromkin (New York, NY: Academic Press), 5–39.

Ohala, M., and Ohala, J. (1972). The problem of aspiration in Hindi phonetics. *Ann. Bull. Res. Inst. Logopedics Phoniatr*. 6, 39–46.

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

Phuong, V. T. (1981). *The acoustic and perceptual nature of tone in Vietnamese* (Ph.D. thesis). Australian National University, Canberra.

Ren, X., and Mok, P. (2021). "Consonantal effects of aspiration on onset F0 in Cantonese," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5.

Samuel, A. G., and Larraza, S. (2015). Does listening to non-native speech impair speech perception? *J. Phon.* 81, 51–71. doi: 10.1016/j.jml.2015.01.003

Schertz, J., Carbonell, K., and Lotto, A. J. (2020). Language specificity in phonetic cue weighting: monolingual and bilingual perception of the stop voicing contrast in English and Spanish. *Phonetica* 77, 186–208. doi: 10.1159/000497278

Schertz, J., Cho, T., Lotto, A., and Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *J. Phon.* 52:183–204. doi: 10.1016/j.wocn.2015.07.003

Schertz, J. L. (2014). *The structure and plasticity of phonetic categories across languages and modalities* (Ph.D. thesis). The University of Arizona, Tucson, AZ.

Shimizu, K. (1994). "F0 in phonation types of initial-stops," in *Proceedings of the 5th Australasian International Conference on Speech Science and Technology, Vol. 2* (Perth), 650–655.

Shultz, A. A., Francis, A. L., and Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *J. Acoust. Soc. Am.* 132, EL95-EL101. doi: 10.1121/1.4736711

Sivula, T., Magnusson, M., and Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv:2008.10296 [stat.ME]*. doi: 10.48550/arxiv.2008.10296

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 1074–1095. doi: 10.1037/0096-1523.7.5.1074

Toscano, J. C., and McMurray, B. (2010). Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci.* 34, 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27:1413–1432. doi: 10.1007/s11222-016-9696-4

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.* 4, 2152–2159. doi: 10.1121/1.406678

Whalen, D. H., and Levitt, A. G. (1995). The universality of intrinsic F_0 of vowels. *J. Phon.* 23, 349–366. doi: 10.1016/S0095-4470(95)80165-0

Winn, M. B. (2020). Manipulation of voice onset time in speech stimuli: a tutorial and flexible Praat script. *J. Acoust. Soc. Am.* 147, 852–866. doi: 10.1121/10.0000692

Xu, B. R., and Mok, P. (2012). "Cross-linguistic perception of intonation by Mandarin and Cantonese listeners," in *Proceedings of Speech Prosody 2012* (Shanghai), 99–102.

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

Yang, J., Zhang, Y., Li, A., and Xu, L. (2017). On the duration of Mandarin tones. *Proc. Interspeech* 2017, 1407–1411. doi: 10.21437/Interspeech.2017-29

Yuan, J., Ryant, N., and Liberman, M. (2014). "Automatic phonetic segmentation in Mandarin Chinese: boundary models, glottal features and tone," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Florence: IEEE), 2539–2543.

Zellou, G. (2017). Individual differences in the production of nasal coarticulation and perceptual compensation. *J. Phon.* 61, 13–29. doi: 10.1016/j.wocn.2016.12.002