Journal of Phonetics 86 (2021) 101049

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

The effect of linguistic experience on perceived vowel duration: Evidence from Taiwan Mandarin speakers

Yu-An Lu [*], Sang-Im Lee-Kim

National Yang Ming Chiao Tung University, Taiwan

ARTICLE INFO

Article history:
Received 27 August 2019
Received in revised form 24 February 2021
Accepted 2 March 2021
Available online 28 March 2021

Keywords:
Perceived vowel duration
Lexical tone
F0 contour
Canonicity effects
Taiwan Mandarin

ABSTRACT

Perceived vowel duration is known to be influenced by many factors, including *f0* height/movement and ones' native phonological system. Using multiple experimental paradigms, this study examined whether native tonal representations and phonetic knowledge of duration associated with different lexical tones may further shape the ways in which vowel duration is perceived. In a perception experiment, Taiwan Mandarin and Korean listeners rated the duration of duration-controlled CV syllables carrying one of the four lexical tones in Mandarin or a reduced T3half (X^{21}). The results showed that perceived vowel duration by Korean listeners, the control group, reflected general perceptual biases: contour tones were rated as longer than level tones, and high-*f0* tones were rated as longer than low-*f0* tones. Taiwan Mandarin listeners, on the other hand, overestimated the duration of vowels carrying T3 (X^{214}) and T3half, despite their short phonetic duration in Taiwan Mandarin, indicating the significance of the canonical representation of the complex T3 contour. A spontaneous imitation experiment further supported the canonicity effect: T3half was again hyperarticulated, produced as longer and with similar *f0* trajectories as T3full, based on its phonological association to T3. Taken together, the findings of the present study suggest that the perception of vowel duration is guided by higher-order phonological knowledge from speakers' linguistic experience as well as by general perceptual biases.

1. Introduction

The perception of vowel duration is known to be influenced by many factors, including properties of the stimulus (e.g., *f0* height, *f0* movement, intensity, and vowel quality) and characteristics of the listener (e.g., native language and musical background). This study investigates whether additional factors—phonetic knowledge of duration associated with different lexical tones (Section 1.4.1) and native canonical tonal representations (Section 1.4.2) —further shape the ways in which vowel duration is perceived by Taiwan Mandarin listeners.

1.1. Characteristics of f0 and perceived vowel duration

Previous studies have established a correlation between perceived vowel duration and *f0* height and movement. For example, Gussenhoven and Zhou (2013) tested speakers of a tone language (Mandarin) as well as a non-tone language (Dutch) on their perceived duration of vowels carried by different *f0* heights (HH [250 Hz] and LL [140 Hz]) and contours (HL [275–140 Hz], LH [140–275 Hz], LHL [140–275–140 Hz], HLH [275–140–275 Hz]) manipulated into different duration steps (230–260–290–320 ms). The results showed that, independent of native language, the listeners perceived vowels with high *f0* as longer than vowels of equal duration with low *f0*. These perceptual results echo those from production studies that have found, cross-linguistically, that vowel duration tends to be inversely related to *f0* height; that is, the lower the tone, the longer the vowel. This is presumably due to the greater physiological difficulty of sustaining vowel duration while maintaining high *f0* (cf. Faytak & Yu, 2011; Gandour, 1977; Šimko, Aalto, Lippus, W□odarczak, & Vainio, 2015).

These results have been taken to support a *perceptual compensation* account in which vowels with high *f0* are perceived as longer than those with low *f0*, all else being equal, to compensate for the shorter duration of high *f0* vowels in production (Gussenhoven, 2004; Gussenhoven & Zhou, 2013). However, a correlation between *f0* height and durational judge-

[*] Corresponding author at: Department of Foreign Languages and Literatures, National Yang Ming Chiao Tung University, 4F, Humanities Building 3, 1001 University Road, Hsinchu 30010, Taiwan.

E-mail address: yuanlu@nctu.edu.tw (Y.-A. Lu).

https://doi.org/10.1016/j.wocn.2021.101049
0095-4470/© 2021 The Author(s). Published by Elsevier Ltd.

ment has also been found in experiments using non-speech stimuli (e.g., Cumming, 2011; Dawson, Aalto, Simko, & Vainio, 2017), rendering the perceptual compensation account less likely. Alternatively, some studies have proposed a central-tendency effect whereby listeners tend to overestimate shorter durations and underestimate longer durations (Jazayeri & Shadlen, 2010; Shi, Church, & Meck, 2013).

Beyond a simple inverse relation between produced duration and perceived duration for vowels with level *f0*, vowels with dynamic *f0* movement present a rather complex picture with respect to perceived duration (Cumming, 2011; Lehiste, 1976; Rosen, 1977; Wang, Lehiste, Chuang, & Darnovsky, 1976). For example, Cumming (2011) tested French and German speakers' relative perceived duration of vowels with different tonal contours: Falling vs. Level, Rising vs. Level, and Complex vs. Level. The results showed that, independent of native language, listeners consistently judged syllables with dynamic *f0* contours as longer than those with static *f0* contours, a perceptual pattern that corresponds *proportionally* to the produced duration. These results were taken to reflect a perceptual skewing that is inherent to the complexity of dynamic *f0* articulation (Yu, 2010).

However, while vowels with complex dynamic contours (e.g., LHL, HLH) are generally produced as longer than those with simple dynamic contours (e.g., LH, HL) (Köhnlein, 2015; Zhang, 2000), the same proportional reflection has not been found in perception (Gussenhoven & Zhou, 2013; Lehiste, 1976). Furthermore, within contour tones, the direction of the *f0* movement may also give rise to a systematic perceptual bias; vowels with rising contours have been found to be perceived as longer than those with falling contours (Dawson et al., 2017; Rosen, 1977; Van Dommelen, 1993; Wang et al., 1976).

1.2. Characteristics of native phonological systems and perceived vowel duration

Alongside the acoustic properties of the stimuli, the prosodic system of the native language has also been shown to have an independent effect on perceived vowel duration (Kinoshita, Behne, & Arai, 2002; Šimko et al., 2015; Takiguchi, Takeyasu, & Giriko, 2010). For example, Šimko et al. (2015) asked Estonian, Swedish, Finnish and Mandarin speakers to discriminate pairs of duration-controlled sounds manipulated to have different tones and intensity. The results revealed that speakers of languages with a quantity system (i.e., Estonian, Swedish and Finnish) were able to judge duration with greater precision. In contrast, Mandarin speakers, whose native language lacks vowel length contrasts, made less precise judgments. Furthermore, among the languages with vowel length contrasts, Estonian and Finnish use *f0* movement to co-signal quantity contrasts, and speakers of these languages demonstrated even higher duration/*f0* sensitivity. The same study also showed that higher *f0* and greater intensity correlated with longer perceived vowel duration, confirming the perceptual tendency outlined above.

While vowel length contrasts have been shown to have an effect on perceived vowel duration, the number of contrastive tones in one's native language does not necessarily increase a listener's sensitivity to vowel duration. Chang and Lu (2016) tested Taiwan Mandarin and Southern Min listeners' perceived duration of duration-controlled vowels with different *f0* heights and contours. Southern Min has a larger tonal inventory than Mandarin (7-tone vs. 4-tone inventory, respectively). The tonal inventory of Southern Min includes two checked tones carried by syllables closed with an unreleased obstruent coda. These tones are characterized by shorter vowel durations compared to unchecked tones carried by longer vowels in open syllables or with nasal codas. Though it was expected that the association between vowel duration and lexical tones may potentially enhance Southern Min listeners' sensitivity to vowel duration, the results showed that the two language groups performed similarly. Consonant inventory, on the other hand, has been shown to have an effect. In the aforementioned study by Gussenhoven and Zhou (2013), it was shown that compared to Dutch listeners, Mandarin listeners perceived vowels with aspirated onsets as longer than those with unaspirated onsets. This was attributed to the phonological difference between the two languages: Mandarin listeners are more likely to interpret aspiration as part of the perceived vowel duration due to the long-lag VOT for the aspirated stops in their native language, as opposed to Dutch listeners whose native language contrasts prevoiced with voiceless unaspirated stops.

The discussion above demonstrates that while *f0* height and shape have an effect on the perceived duration of vowels (Section 1.1), the native phonological system may also influence vowel duration judgments (Section 1.2). This raises the question of how the perception of vowel duration by speakers of a tone language, who use *f0* contours to signal phonological contrasts, may be affected by varying *f0* vowel trajectories. The aforementioned study by Gussenhoven and Zhou (2013) was an attempt to address this question. In their study, Mandarin and Dutch listeners judged the duration of vowels in duration-controlled stimuli with different *f0* contours resembling Mandarin tones. The results showed a negative correlation between the perceived duration judgments and the actual durations of different Mandarin tones taken from Whalen and Xu (1992). This appears to be in accordance with the perceptual compensation account whereby tones produced with long durations are perceived as shorter and tones produced with short durations are perceived as longer.

However, Gussenhoven and Zhou's results need to be interpreted with caution for several reasons. First, their stimuli were produced by a native Russian speaker, and the *f0* contours of the stimuli were not good mappings of Mandarin tones. In particular, some of the *f0* contours (e.g., LHL [140-275-140 Hz]) are not attested in Mandarin Chinese and would likely sound completely foreign to Mandarin listeners. These unfamiliar stimuli may have impeded Mandarin listeners from utilizing their linguistic knowledge of the lexical tones. The fact that the Dutch listeners behaved similarly indicates that the particular experimental design was suitable for testing a shared mechanism such as perceptual compensation, independent of language background. The methodological constraints of Gussenhoven and Zhou's study necessitate experimentation using stimuli that closely reflect the phonetic properties of the lexical tones such that participants could be encouraged to perform in their native-language mode. To this end, the current study employs stimuli with tones modeled on the four lexical tones in Mandarin to examine how *canonical tonal representations* and the *redundant phonetic knowledge of duration associated with different tone contours* further shape the ways in which vowel duration is perceived. Mandarin Chinese was chosen for its different tonal contrasts (i.e., level, simple contour and complex contour) associated with different temporal cues and its lack of vowel length contrasts.

1.3. An overview of lexical tones in Mandarin Chinese

Standard Mandarin Chinese has four phonemic tones, high-level Tone 1 [X^{55}], rising Tone 2 [X^{35}], falling-rising Tone 3 [X^{214}], and falling Tone 4 [X^{51}]. The tone numbers ranging from 1 to 5 here indicate relative *f0*—the higher the tone number, the higher the *f0* (Chao, 1968: 25-30). The canonical realizations of the tones are often found in the last syllable of a word as in Fig. 1 (Chao, 1968; Duanmu, 2007; Lin, 2007; Xu, 1997).[1] In addition to the canonical *f0* trajectories of the four lexical tones, T3 [X^{214}] has a reduced variant, T3half [X^{21}], when followed by a tone other than itself (Yip, 2002; Zhang & Lai, 2010). T3half is short in duration and low in *f0*, and its *f0*-trajectory no longer forms a contour shape, as shown in the first syllable in Fig. 1. Another variant of T3 is T2, arising via tone sandhi when a T3 occurs before another T3. This variant of T3 is excluded from the discussion since previous studies generally agree that the tonal contours of the sandhi-ed T2 and the lexical T2 are indistinguishable (Chien, Sereno, & Zhang, 2017; Myers & Tsay, 2003; Zhang & Lai, 2010).

This gives rise to Mandarin tones contrasting up to three types on the surface: level (T1, T3half as an allotone of T3), simple dynamic contour (T2, T4), and complex dynamic contour (T3) tones. Based on these particular tonal contours, when controlling for phonetic duration, the following predictions can be made. If perceived vowel duration is *language-independent*, listeners' perception should follow the *general perceptual biases* outlined in Section 1.1. Each bias makes a specific prediction about duration-controlled vowels carried by different Mandarin tones.

(1) **General perceptual biases**

Bias 1 (dynamic > static)

Vowels with a dynamic *f0* should be perceived as longer than those with a static *f0*:
T2 [X^{35}], T4 [X^{51}], T3 [X^{214}] > T1 [X^{55}], T3half [X^{21}]

Bias 2 (rising > falling)

Vowels with a rising contour should be perceived as longer than those with a falling contour:
T2 [X^{35}] > T4 [X^{51}]

Bias 3 (high > low)

Vowels with a high *f0* should be perceived as longer than those with a low *f0*:
T1 [X^{55}] > T3half [X^{21}]

In the present study, this hypothesis was tested against Seoul Korean listeners without experience with any tone language. Prosodically, Seoul Korean is neither tonal nor stressed, but is considered a phrasal-accent language (Jun, 2005). Although Korean is traditionally known to be quantity-sensitive, vowel length contrasts have been shown to be levelled for young Seoul Korean speakers (Kwon, 2003), the group chosen for participation in this study. We predict that Korean listeners' perceived vowel duration should follow these general perceptual tendencies.

1.4. Effects of higher-order linguistic knowledge on perceived vowel duration beyond psycho-acoustic processing

In addition to the general biases drawn from the acoustic properties of *f0* trajectories, we explore the role of higher-order linguistic knowledge on perceived vowel duration. Specifically, we test the two following linguistic factors: i) phonetic knowledge of the durational differences associated with lexical tone in one's native language and ii) canonical forms and phonological categories (i.e., T3 and its variant T3half).

1.4.1. Phonetic knowledge of varying vowel durations associated with lexical tones

Although Mandarin lacks phonemic vowel length contrasts, lexical tones come with different phonetic durations as exemplified by the relative durations illustrated in Fig. 1. In a corpus study examining Mandarin syllable durations with primarily Mainland Mandarin speakers, Wu and Kenstowicz (2015) showed that Mandarin lexical tones followed a scale from longest to shortest: T3 (X^{214}, *M* = 407 ms) > T2 (X^{35}, *M* = 364 ms), T1 (X^{55}, *M* = 333 ms) > T4 (X^{51}, *M* = 286 ms). Notice that the durational properties associated with the lexical tones are well-aligned with the cross-linguistic patterns discussed earlier. T3, the most complex tone in terms of *f0* movement, is the longest in production while the falling T4 is the shortest. Further, the rising T2 is produced as longer than the falling T4 (see similar findings in Ho (1976) and Xu (1997)).

However, a different duration scale is observed for the lexical tones in Taiwan Mandarin, the target language in this study. The corpus study reported in the next section shows that in Taiwan Mandarin, T3 is shorter in production than T2 and T1 when it is spoken in isolation and is comparable to T4 when spoken in a frame sentence, yielding a significant demotion of T3 in the relative duration scale: T2 > T1 > T3 > T4.

These varying durational patterns in Mandarin lexical tones generate another hypothesis for Taiwan Mandarin listeners' perceived vowel duration. If the perceived duration is, to some extent, informed by the *phonetic knowledge* associated with different lexical tones in their native language, then Taiwan Mandarin listeners' perceived duration should follow that of the produced duration of the different tones, as outlined below:

(2) **Phonetic bias**: Perceived durations mimic produced durations
T2 [X^{35}] > T1 [X^{55}] > T3 [X^{214}] > T4 [X^{51}]

[1] Our discussion here excludes the neutral tone (T0 or T5), which only occurs in morphologically weak positions and has varying tone values depending on preceding tones.

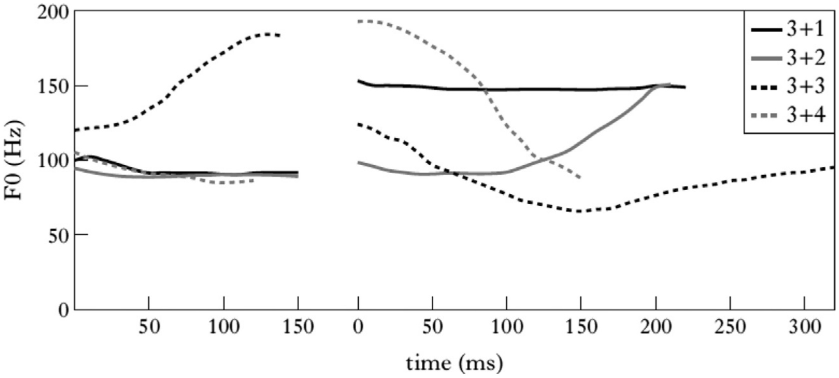

Fig. 1. Representative f0 tracks for the tone sandhi in Mandarin (taken from Zhang & Lai, 2010: 163).

This hypothesis differs from the sets of General Perceptual Biases outlined in (1), which predict that T4, a contour tone, should be perceived as longer than T1 and T3half, the two level tones. Moreover, while hypotheses based on General Perceptual Biases make no specific predictions about differences between simple and complex contours, the phonetic bias hypothesis in (2) is unequivocal about their relative scale.

1.4.2. Canonical forms and phonological categories

Previous studies have shown an effect of canonicity in speech processing (Chuang, 2017; Fon, Hung, Huang, & Hsu, 2011; Sumner & Samuel, 2005, 2009). For example, in a series of experiments using semantic priming and lexical decision tasks, Sumner and Samuel (2005) found that the target word *music* was immediately primed by the semantically related word *flute* when articulated with any of the three variants of the final /t/: fully aspirated [t^{h}], unreleased [ʔt˺] and glottalized [ʔ], while a contrastive phoneme /s/, as in [flus], had no facilitation effect on processing the target word *music*. More significantly, when the primes and targets were presented in two different blocks, long-term priming was found only for the fully aspirated *canonical* [t^{h}], despite the fact that the unreleased [ʔt˺] is the most frequent variant. The fully aspirated [t^{h}] is considered canonical because it is the unreduced ideal form that occurs in word-initial and stress-syllable initial positions (Jaeger, 1980; Sumner & Samuel, 2005).

It is worth noting that canonical forms do not necessarily refer to acoustically salient variants. Sumner and Samuel (2009), for example, examined cross-dialectal spoken word recognition comparing speakers from the rhotic American English variety with those from the non-rhotic NYC variety. Interestingly, the rhotic primes (e.g., [beɪkɚ] 'baker') facilitated the target recognition to a greater degree than the non-rhotic variants (e.g., [beɪkə]) for the *r*-less NYC speakers, as well as speakers from the standard variety. Since neither form is more salient than the other, the authors concluded that the canonicity associated with the rhotic form is the primary reason for the better recognition in word processing (Sumner, Kim, King, & McGowan, 2014; Sumner & Samuel, 2009). These findings are suggestive of an effect of prototypical or canonical forms that are stored in the mental representation.

Along the same lines, in a series of priming experiments, Chuang (2017) investigated immediate word recognition involving canonical vs. non-canonical variant primes for sibilants in Taiwan Mandarin. Due to frequent contact with Taiwanese Southern Min, which lacks retroflex sibilants, Taiwan Mandarin speakers often merge alveolar-retroflex sibilants through deretroflexion of the retroflex sibilants (e.g., /ʂ/→[s]). Despite the frequent appearance of the variant forms, however, canonical primes (i.e., fully retroflexed sibilants) elicited strong facilitatory effects in word processing, namely fewer errors and faster response times, for the target sibilants. These examples demonstrate that the abstract canonical representations have a positive effect on word processing.

Building upon these previous works, the current study investigates whether the impact of canonicity could be extended to the perception of suprasegmental features. As mentioned in Section 1.3, the reduced variant T3half [X^{21}] (occurring when T3 is followed by a tone other than itself) is the most commonly attested form of all T3 variants. In fact, the canonical T3 with its full concave contour is limited to the final position and in isolation (e.g., Duanmu, 2007; Lin, 2007; Qu, 2013; Van de Weijer & Sloos, 2014). However, T3 [X^{214}] is canonically represented as a falling-rising contour in *Zhuyin*, an alphabetic representation of Mandarin used in Taiwan, (a diacritic /ˇ/ next to the phonetic symbols, e.g., /□□ˇ/ [ni^{214}] 'you'), and in the prescriptive grammar.[2] Moreover, when Taiwan Mandarin speakers are specifically asked to produce T3full, a full concave contour can be elicited spontaneously (as in the base stimuli created for our perception study, see Fig. 4). If the canonical tonal representation does influence speech perception, the complex dynamicity in the contour of T3full is likely to lead to an overestimation of the vowel duration. Further, if Taiwan Mandarin listeners' perceived vowel duration is, to some extent, guided by higher-order linguistic knowledge (i.e., the relationship between T3 variants) both T3half and T3full are likely to be overestimated. This leads to the canonicity bias as stated in (3).

[2] In Zhuyin (also called Bopomofo or Mandarin Phonetic Symbols), as well as in Pinyin, the system employed in China, T3 is marked as ˇ, indicating its complex falling-rising contour.

(3) **Canonicity Bias:** Overestimation of the complex dynamic contour T3full [X^{214}] and its variant T3half [X^{21}].

We chose the variety of Mandarin spoken in Taiwan (hereafter TM) as a test case for the following reasons. Although it has been shown that the temporal differences of different Mandarin lexical tones follow a particular scale from longest to shortest (i.e., T3 > T2, T1 > T4) in the speech of Mainland Mandarin speakers (cf. Section 1.4.1), the corpus data collected in the current study show that, in TM, T3 is produced as shorter than T1 and T2 when spoken in isolation and is produced with a duration comparable to T4 when spoken in a frame sentence. In other words, while the most complex tone is produced as the longest in other varieties of Mandarin, this is not the case in TM. This creates a conflict between the canonical tonal representation and the corresponding durational difference, giving us a unique chance to tease apart the hypotheses being tested. Specifically, T3 and its variant T3half are predicted to be overestimated in perception by virtue of the Canonicity Bias, which cannot be tested with speakers of Mainland Mandarin due to the confounding factor that the most complex tone is also the longest in phonetic duration.

1.5. Summary

The discussion thus far has led us to make the following predictions for TM speakers' perceived vowel duration. If perceived vowel duration is language-independent, TM listeners' perceived duration should follow the general perceptual biases and show comparable patterns with Korean listeners. That is, listeners of both languages would perceive vowels with dynamic *f0* as longer than those with static *f0* (T2, T3, T4 > T1, T3half), vowels with a rising contour as longer than those with a falling contour (T2 > T4), and vowels with high *f0* as longer than those with low *f0* (T1 > T3half). On the other hand, if perceived duration is informed by TM listeners' phonetic knowledge associated with different tones from linguistic experience, their perceived vowel duration should reflect the produced duration (T2 > T1 > T3 > T4). Alternatively, if TM listeners' perceived vowel duration is, to some extent, guided by canonical tonal representations, T3 is likely to be overestimated due to its complex dynamicity in contour, and the lexical link to the T3 category may further lead to the overestimation of T3half as well.

These hypotheses were tested through a series of experiments in the present study. We first conducted a corpus study to establish an empirical foundation for the temporal differences associated with different lexical tones in TM. The results of this study were expected to demonstrate a language-specific Phonetic Bias (2) arising from the TM listeners' experience with their native language. We then examined the validity of the aforementioned hypotheses explicitly through a cross-linguistic perception study in which perceptual patterns of native speakers of TM were compared with those of Korean listeners without any experience in tone languages. To strengthen the empirical grounding of the findings of our perception study, we conducted a spontaneous imitation experiment to probe how listeners' phonological knowledge of lexical tones may affect their production of those tones. Output from imitation tasks has been shown to reflect abstract phonological representations as well as fine-grained phonetic details in the stimuli (H. Kwon, 2019; Mitterer & Ernestus, 2008; Nielsen, 2011; Scarborough, Strickler, & Nielsen, 2020). We thus would expect to see a solid connection between perceptual biases and the native phonological system.

2. Corpus study: Lexical tones and durations in TM syllables

Although it has been widely noted that TM and Mainland Mandarin (or Putonghua) differ in terms of the phonetic realization of tones (e.g., Fon & Chiang, 1999; Deng, Shi, & Lu, 2006; Kubler, 1985), there is no sizable database to refer to the phonetic patterns in TM. Therefore, we began our investigation with a corpus study to provide an empirical foundation for the temporal differences associated with different lexical tones in TM.

2.1. Methodology

2.1.1. Participants

10 TM speakers (8 female, 2 male; aged 20–26; $M = 21.3$) were recruited to provide production data. None of the participants reported hearing or speaking deficiencies. All participants were compensated monetarily for their time.

2.1.2. Materials

A list of 112 Mandarin Chinese monosyllabic words was compiled.[3] These words were balanced for tone (T1, T2, T3, T4) and syllable type (CV, CGV, CVN, CGVN) and were comparable in terms of segments (vowels and consonants) to avoid intrinsic durational differences across different segments. To avoid any frequency effects (i.e., the tendency of more frequent words being produced as shorter (cf. Bybee, 2003; Gahl, 2008; Pierrehumbert, 2001)), word frequencies were balanced across the four tones using the *Taiwan Mandarin Conversational Corpus* (Tseng, 2013).

2.1.3. Procedure

The participants were recorded individually in a sound attenuated booth using a Marantz PMD661A recorder at a sampling rate of 44.1 kHz, 16 bits. An AKG P220 large-diaphragm condenser microphone was connected to the recorder and placed on a stand, facing the participant's mouth approximately 20 cm away. The 112 words were presented twice in random orders using E-Prime (Schneider, Eschman, & Zuccolotto, 2002) in two contexts, one in isolation and the other in a frame sentence (/$t^h a^{55}$ $tso\eta^{214}$ $ʂi^{51}$ pa^{214} _____ tu^{35} $ts^h uo^{51}$/, 'He always pronounces ____wrong') presented in Chinese characters in two separate blocks. The resulting corpus contained 4,480 words (=10 speakers × 112 words × 2 contexts × 2 repetitions). The participants were instructed verbally as well as with written instructions on the computer screen to read the words and sentences presented to them at a comfortable speaking rate. Four practice trials were presented before recording began to familiarize participants with the task. The total duration of the procedure was around 20 min.

[3] All stimuli are available at https://osf.io/kcm2s/?view_only=e63ccbd9196f4b97a9caf0ce11f7c078.

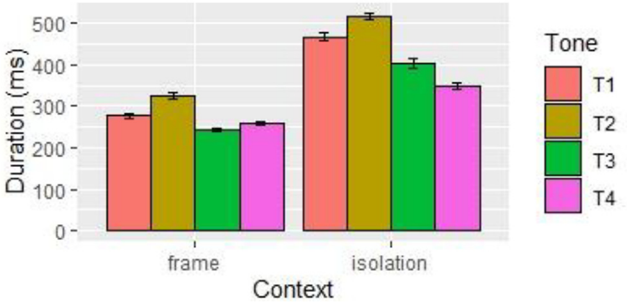

Fig. 2. Mean duration of target words by tone with error bars representing 95% confidence intervals.

The target syllables were first labeled in Praat (Boersma & Weenink, 2017) to measure the durations associated with different tones. Obstruent onsets were further annotated out to obtain continuous *f0* contours. ProsodyPro (Xu, 2013) was employed to perform time-normalization of *f0* converted to a semitone scale.

2.2. Results

Fig. 2 shows the mean duration of the words as a function of tone in the two contexts.[4] In general, words produced in a frame sentence (*M* = 275.82 ms) were shorter than those produced in isolation (*M* = 434.14 ms). From the graph, it is evident that T2 is the longest followed by T1 in both reading contexts. This presents a sharp contrast with the tone-duration association in Mainland Mandarin wherein T3 is the longest.

To assess the statistical significance of these differences, a mixed-effects linear regression analysis was run in R using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015), and associated *p*-values were obtained using the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2016). The dependent variable was the word duration converted into *z*-scores for each speaker to accommodate individual differences in speaking rate. The model included fixed effects for Context (frame, isolation), Tone (T1, T2, T3, T4), and their interaction. The model also included the random intercept and slope for Participant and the random intercept for Item. With isolation context and T1 as baselines, the model showed that T1 was shorter than T2 (β = 0.43, p < .01) but longer than T3 and T4 (β = −0.57, p < .01 and β = −0.98, p < .0001, respectively) when produced in isolation. A separate model with T3 as the baseline further confirmed that T3 was produced as significantly longer than T4 (β = −0.41, p < .01).

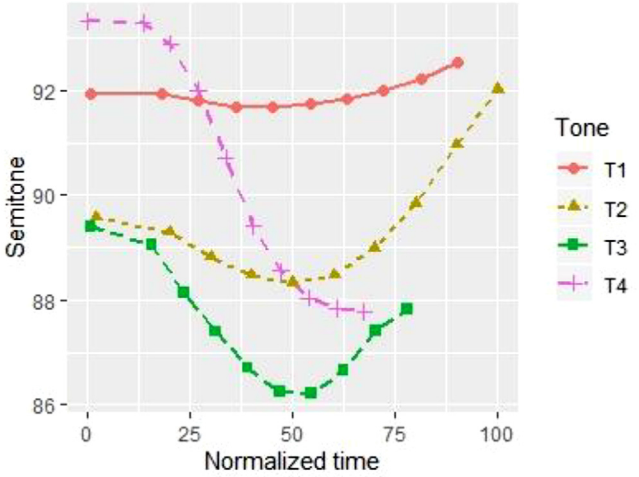

Fig. 3. Time normalized semitones from the isolation context as a function of Tone.

Further, significant Tone-Context interactions were found for T1-T3 (β = 0.27, p < .0001) and T1-T4 (β = 0.81, p < .0001) but not for T1-T2 (β = −0.01, p = .85). This indicates that the durational patterns of T3/T4 were different depending on the reading context. We further confirmed the durational differences among the different lexical tones in the frame condition by setting the frame context and T3 as baselines. The results showed that T2 was still the longest tone (β = 0.72, p < .0001), followed by T1, which was longer than T3 at a marginally significant level (β = 0.30, p = .08). T3 and T4 were not statistically different (β = 0.13, p = .36).

The statistical analyses led to the establishment of the durational differences across the different lexical tones, as summarized in Table (4) below. The statistical models are summarized in (5); for the formulas that generated these results, see Appendix 1.[5] The duration of T3 was significantly longer than that of T4 in the isolation condition; however, this durational difference diminished in the frame condition, suggesting that T3 was realized as its reduced variant, T3half, in the frame context.

(4) Durational difference in TM lexical tone with mean duration in parentheses

Context	Durational differences
Isolation	T2 (*M* = 517.44) > T1 (*M* = 467.00) > T3 (*M* = 403.74) > T4 (*M* = 348.39)
Frame	T2 (*M* = 324.82), T1 (*M* = 276.81) > T3half (*M* = 243.25) T3half = T4 (*M* = 258.40)

[4] All data are available at https://osf.io/kcm2s/?view_only=e63ccbd9196f4b97a9caf0ce11f7c078.

[5] The purpose of the corpus study was to establish the relative order between produced durations of each tone and thus the post-hoc models were treated as planned comparisons. No adjustments were made to the *p*-values.

(5) Summary of fixed effects for the corpus study

Isolation/T1 as baselines					Isolation/T3 as baselines					Frame/T3 as baselines				
R^2 = 0.76					R^2 = 0.76					R^2 = 0.76				
Predictor	B	SE	t	p		B	SE	t	P		B	SE	t	p
(Intercept)	0.88	0.09	9.67	<0.0001	(Intercept)	0.31	0.17	1.84	0.09	(Intercept)	−0.89	0.09	−9.57	<0.0001
T2	0.43	0.12	3.44	<0.01	T1	0.57	0.17	3.44	<0.01	T1	0.30	0.17	1.81	0.08
T3	−0.57	0.17	−3.44	<0.01	T2	1.00	0.18	5.61	<0.0001	T2	0.72	0.18	4.05	<0.0001
T4	−0.98	0.12	−7.99	<0.0001	T4	−0.41	0.14	−2.90	<0.01	T4	0.13	0.14	0.93	0.36
Frame	−1.47	0.14	−10.28	<0.001	Frame	−1.19	0.14	−8.39	<0.0001	Isolation	1.19	0.14	8.40	<0.0001
T2:Frame	−0.01	0.04	−0.19	0.85	T1:Frame	−0.27	0.04	−6.55	<0.0001	T1:Isolation	0.27	0.04	6.55	<0.0001
T3:Frame	0.27	0.04	6.55	<0.0001	T2:Frame	−0.28	0.04	−6.74	<0.0001	T2:Isolation	0.28	0.04	6.74	<0.0001
T4:Frame	0.81	0.04	19.64	<0.0001	T4:Frame	0.54	0.04	13.09	<0.0001	T4:Isolation	−0.54	0.04	−13.09	<0.0001

We further verified the realization of tonal contours of the words produced in isolation by the TM speakers. In order to compare tonal contours across speakers and tones, time and *f0* were normalized. The results are shown in Fig. 3.

The results showed that while the contours of T1, T2 and T4 were comparable between TM speakers and Mainland Mandarin speakers (*cf*. Fig. 1), the contours of T3 between the two dialects were considerably different. We observed that the final rise of T3 fell short of its high target in TM speakers' production, consistent with previous studies (Fon & Chiang, 1999; Kubler, 1985). That is, while a canonical T3 produced in citation in Mainland Mandarin ends with a much higher *f0*, as indicated by X^{214} in Chao's letter system, in TM speakers' production of T3, *f0* at the end is even lower than it is at the beginning and would thus be best transcribed as X^{212}.

2.3. Discussion

While previous studies on Mandarin Chinese have found that T3 is generally produced as the longest (Ho, 1976; Wu & Kenstowicz, 2015; Xu, 1997), in TM speakers' tone production, T2 was the longest, while T3full was generally short and T3half was comparable to T4, the shortest tone.[6] The shorter duration of T3, even when produced in isolation, may have resulted from the absence of the final rise or compression of the rise in duration observed in Fon and Chiang (1999) and Kubler (1985). Taken together, unlike in other varieties of Mandarin, the duration of the final rise in T3full is significantly compressed in TM, which in turn results in an overall shorter duration in production. These results provide an empirical foundation to examine the three hypotheses outlined in (1)-(3) in the perceived vowel duration experiment presented in the next section. The prediction of the **Phonetic bias hypothesis** (outlined in (2)) is based on the assumption that one's phonetic experience influences the perceived vowel duration. Specifically, T3-full, despite its complex *f0* trajectory, is predicted to be demoted on the durational scale due to its shorter phonetic duration in TM speakers' production. The scale in (2) is reproduced below in (6) with the addition of T3half.

(6) **Phonetic bias hypothesis:** Produced durations faithfully reflected in perceived durations
T2 [X^{35}] > T1 [X^{55}] > T3 [X^{214}] > T4 [X^{51}] ≈ T3half [X^{21}]

3. Perceived vowel duration experiment

This section presents an experiment designed to examine the perceptual patterns of TM listeners in comparison with those of Korean listeners without any experience in tone languages.

3.1. Methodology

3.1.1. Participants

20 TM speakers (10 female, 10 male; aged 19–36; *M* = 22.9) were recruited at National Chiao Tung University to serve as the target group. 20 Korean speakers (12 female, 8 male; aged 19–34; *M* = 24.7) recruited at Seoul National University served as the non-tone language baseline group. None of the Korean participants had studied Mandarin nor did they speak Korean dialects with pitch accents. None of the participants reported hearing or speaking deficiencies. All participants were compensated monetarily for their time.

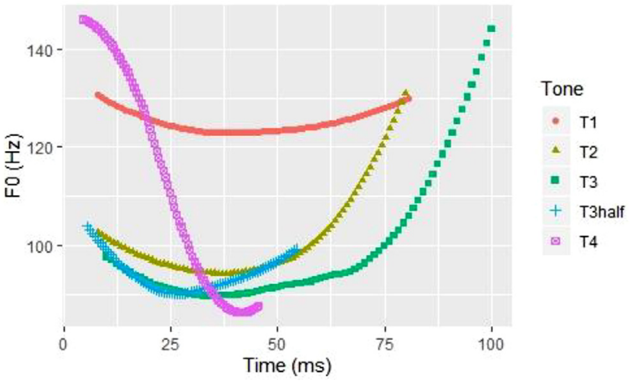

Fig. 4. Mean f0 trajectories and durations of the naturally produced tokens.

[6] This pattern was also found in the TM speaker's data in Wu and Kenstowicz (2015: 91).

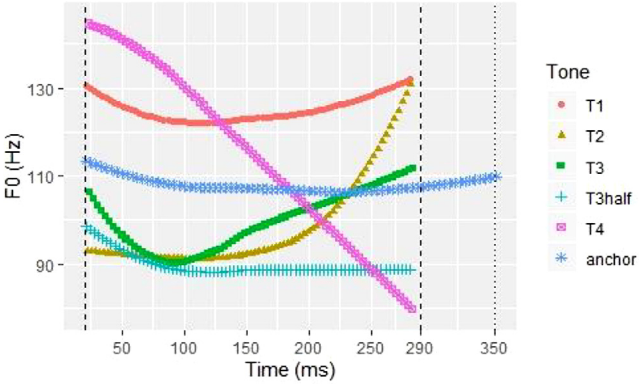

Fig. 5. Example stimuli [pa] resynthesized to 290 ms and the anchor stimuli set to 350 ms with a mid-level tone.

3.1.2. Stimuli

Four CV syllables, [pa], [pi], [ta], and [ti], were selected as the target syllables. Unaspirated consonants were used as speakers might perceive aspiration as part of the vowel duration (Chang & Lu, 2016; Gussenhoven & Zhou, 2013). The two vowels /i/ and /a/ were used to add variation to the stimuli. All syllable-tone combinations were well-formed real words in Mandarin to avoid any possible bias against nonwords. A TM male speaker with some phonetic training produced the tokens in isolation. The speaker was asked to produce T3full and T3half separately. Fig. 4 illustrates the *f0* trajectories and durations of the recorded tokens. As shown in this figure, T3full was somewhat hyper-articulated with the *f0* of the end of the tone being even higher than that of T2 (X^{35}). But this precisely reflects the canonical representation of T3, consistent with the high *f0* ending and longer duration as indicated by X^{214} in Chao's tone letter system.[7]

These tokens were then resynthesized into a five-step duration continuum, 290–320-350–380-410 ms, falling well within the duration range of Mandarin syllables, using the Pitch Synchronous Overlap and Add (PSOLA) algorithm in Praat (Boersma & Weenink, 2017). Fig. 5 shows an example of the stimuli resynthesized to 290 ms. A fixed stimulus [pa] set to 350 ms in a mid-level tone was also resynthesized to serve as an anchor. Note that the tonal trajectories in Fig. 5 appear to be different from those in Fig. 4. The differences stem from the fact that Fig. 4 plotted the trajectories and durations of all the naturally produced tokens while Fig. 5 only plotted those of one representative token, [pa].

3.1.3. Procedure

The procedure closely followed that in Gussenhoven & Zhou (2013). The 100 resynthesized stimuli (4 syllables [pa, pi, ta, ti] × 5 tones [T1, T2, T3, T3half, T4] × 5 duration steps [290–320-350–380-410 ms]) were presented twice in two blocks using E-Prime (Schneider et al., 2002). The 200 trials were randomized for each participant and presented in an AX task in which "A" was the anchor stimulus ([pa] fixed at 350 ms in a mid-level tone) and "X" was the target stimulus. The interstimulus interval (ISI) was set as 800 ms. Participants were instructed verbally in their respective languages as well as with written instructions on the computer screen. They were asked to listen to each pair of sounds and judge the relative duration of the target stimulus compared to the anchor stimulus. The judgments were made on a 7-point scale, with 1 indicating that the target was much shorter than the anchor, 4 that the target and the anchor were the same duration, and 7 that the target was much longer than the anchor. Participants used the number keys on a keyboard connected to a computer.

Six practice trials were presented before the experiment to familiarize participants with the task. These trials contained the experimental stimuli in either the longest duration step (410 ms) or the shortest (290 ms) randomly chosen from the four syllables and five tone combinations. The experiment was conducted in sound-attenuated booths using high-quality headphones in two separate locations, one at National Chiao Tung University in Taiwan for TM listeners and the other at Seoul National University in Korea for Korean listeners. The total duration of the experiment was around 20 min.

3.2. Results

The aggregated results are shown in Fig. 6, with the five-step resynthesized vowel durations plotted along the *x*-axis, and the listeners' perceived vowel duration judgments on the 1–7 scale plotted on the *y*-axis. TM listeners' judgments are presented on the left, Korean on the right. In general, both TM and Korean listeners were sensitive to the duration manipulation of the stimuli, as indicated by the proportional relation between perceived duration and manipulated vowel duration. Participants were also sensitive to lexical tones. T2, T3 and T4 (i.e., the three contour tones) were generally judged as longer than T1 and T3half, the two level tones, but with different orderings of tones between the two language groups.

To interpret the results, a linear mixed-effects regression model was fitted to the data in R.[8] The dependent variable was the participants' judgments of vowel durations converted to *z*-scores for each speaker. The model included fixed effects for Group (2 levels: TM = −1 vs. Korean = 1), Tone (5 levels: T1, T2, T3, T3half, T4), Duration (5 steps, converted to *z*-scores centered around 350 ms, the duration of the anchor stimulus, to avoid extrapolation below 290), and Vowel (2 levels: /a/= −1 vs. /i/= 1). The binary variables Group and Vowel were contrast coded so the sum of the weight of each level would be 0 (Davis, 2010). The model included random intercepts for Participant as well as by-participant random slopes for the fixed factors. In addition, the model included interaction terms for Vowel and Tone, Vowel and Duration, and Group, Tone, and Duration. The results of the statistical model are summarized in (7); for the formulas that generated these results, see Appendix 2.

[7] One reviewer raised the concern that the stimuli used for T3full is marked in TM and one would hardly find such realizations unless in a pedagogical setting. However, these tokens were judged by nine native TM speakers to be good representations of T3 prior to the perception experiment. In fact, six out of the ten speakers (ID numbers 1, 4, 7, 8, 9, 10) who participated in the corpus study (Section 2) produced instances of T3 in isolation with high-*f0* ending contours. Although in TM the fully concave T3 is not as frequent as its reduced variant, it is well attested in both perception and production.

[8] The linear models treated the duration judgments as a ratio scale and the steps along the scale as equal in magnitude. We first checked the distribution of the *z*-scored duration ratings to confirm that there was an approximately linear function of mean ratings with stimulus durations and that the steps along the rating scale were uniform in size. We further verified that the residuals of the linear mixed effects model were close to normal.

The statistical model fit showed that the effect of Vowel was significant ($\beta = -0.04$, $p = .03$), driven by the generally longer perceived duration of [a] ($M = 4.29$) compared to [i] ($M = 4.23$). However, as Vowel-Duration ($\beta = -0.02$, $p = .04$) and Vowel-Tone interactions (Vowel-T3half: $\beta = 0.08$, p = .002; Vowel-T4: $\beta = 0.21$, $p < .0001$) were shown to be significant, we further inspected the stimuli used in this experiment. We found that there was a more obvious rise in the T3half carried by the vowel [i] than in that carried by the vowel [a]. The rise in T2 was also slightly more pronounced for [i] than for [a]. As such, these interactions may be a reflection of the variation in the phonetic realization of the tones across the different syllables in the stimuli.

We now turn to the primary focus of the study and examine the potential biases at play in the duration judgements, starting with the General Perceptual Bias 1 (dynamic > static) hypothesis. With T1 as the baseline, the model showed that T2 ($M = 4.6$, $\beta = 0.61$, $p < .0001$), T3 ($M = 4.63$, $\beta = 0.64$, $p < .0001$) and T4 ($M = 4.28$, $\beta = 0.31$, $p = .002$) were perceived as longer than T1 ($M = 3.89$) at the baseline of 350 ms, while the perceived durations between T3half ($M = 3.92$) and T1 were indistinguishable ($\beta = 0.01$, $p = .93$). The lack of interactions between T2/T3/T4 and Group ($\beta = -0.09$, $p = .23$; $\beta = -0.26$, $p = 08$; $\beta = -0.10$, $p = .29$) suggest that the dynamic tones (T2, T3, T4) were, in general, judged as longer than the level tones (T1, T3half), independent of language background. The results directly oppose the prediction generated by the Phonetic Bias hypothesis (6), in which longer produced duration was predicted to be perceived as longer. In TM, high-level T1 is generally produced with a longer duration than contour T3 and T4; however, both T3 and T4 were perceived as longer than T1.

We did find a significant T3half-Group interaction ($\beta = -0.29$, $p = .02$), indicating a group difference in the two level tones (T1 baseline and T3half): high-level T1 was perceived as longer by the Korean participants while TM participants perceived low-level T3half as longer. This shows that the results of the Korean listeners were in line with General Perceptual Bias 3 (high > low), but those of the TM listeners were not.

To test General Perceptual Bias 2 (rising > falling), we fitted a model with T2 as the baseline. The results showed that T2 was perceived as longer than T4 at the baseline 350 ms ($\beta = -0.30$, $p = .003$), and the lack of Group effect ($\beta = 0.06$, $p = 0.33$) and T4-Group interaction ($\beta = -0.01$, $p = .92$) suggests that this tendency was true for participants in both language groups. The statistical models are summarized in (7) and the aggregated data of the perceived vowel duration carried by different tones are ordered in (8) for each language group. Contour tones are indicated in italics, T3 categories in bold.

(7) Summary of fixed effects for the perceived vowel duration experiment

	T1 as baseline					T2 as baseline			
	$R^2 = 0.43$					$R^2 = 0.43$			
Predictor	B	SE	t	p		B	SE	t	p
(Intercept)	−0.32	0.07	−4.42	<0.0001	(Intercept)	0.30	0.06	5.30	<0.001
Vowel	−0.04	0.02	−2.13	0.033	Vowel	0.00	0.02	−0.01	0.989
Duration	0.23	0.03	8.03	<0.0001	Duration	0.44	0.03	15.55	<0.001
T2	0.61	0.07	8.24	<0.0001	T1	−0.61	0.07	−8.25	<0.001
T3	0.64	0.14	4.54	<0.0001	T3	0.03	0.12	0.24	0.816
T3half	0.01	0.12	0.08	0.933	T3half	−0.60	0.11	−5.73	<0.001
T4	0.31	0.09	3.38	0.002	T4	−0.30	0.09	−3.20	0.003
Group	0.15	0.07	2.06	0.047	Group	0.06	0.06	0.99	0.328
Vowel:Duration	−0.02	0.01	−2.02	0.044	Vowel:Duration	−0.02	0.01	−2.02	0.044
Vowel:T2	0.04	0.03	1.50	0.134	Vowel:T1	−0.04	0.03	−1.50	0.134
Vowel:T3	−0.01	0.03	−0.19	0.852	Vowel:T3	−0.05	0.03	−1.68	0.092
Vowel:T3half	0.08	0.03	3.16	0.002	Vowel:T3half	0.04	0.03	1.66	0.098
Vowel:T4	0.21	0.03	7.65	<0.0001	Vowel:T4	0.17	0.03	6.15	<0.001
Duration:T2	0.22	0.03	7.99	<0.0001	Duration:T1	−0.22	0.03	−7.99	<0.001
Duration:T3	0.11	0.03	3.96	<0.0001	Duration:T3	−0.11	0.03	−4.03	<0.001
Duration:T3half	0.19	0.03	6.93	<0.0001	Duration:T3half	−0.03	0.03	−1.07	0.287
Duration:T4	0.15	0.03	5.60	<0.0001	Duration:T4	−0.06	0.03	−2.39	0.017
T2:Group	−0.09	0.07	−1.22	0.229	T1:Group	0.09	0.07	1.23	0.228
T3:Group	−0.26	0.14	−1.82	0.076	T3:Group	−0.17	0.12	−1.43	0.162
T3half:Group	−0.29	0.12	−2.38	0.023	T3half:Group	−0.20	0.11	−1.85	0.072
T4:Group	−0.10	0.09	−1.078	0.288	T4:Group	−0.01	0.09	−0.097	0.924
Duration:Group	−0.01	0.03	−0.5	0.618	Duration:Group	0.06	0.03	2.12	0.037
Duration:T2:Group	0.07	0.03	2.78	0.005	Duration:T1:Group	−0.07	0.03	−2.78	0.005
Duration:T3:Group	0.03	0.03	1.287	0.198	Duration:T3:Group	−0.04	0.03	−1.491	0.136
Duration:T3half:Group	0.08	0.03	2.884	0.004	Duration:T3half:Group	0.00	0.03	0.105	0.916
Duration:T4:Group	0.10	0.03	3.675	<0.0001	Duration:T4:Group	0.02	0.03	0.894	0.371

(8) Results summary for perceived vowel duration from longest to shortest

Both Groups		*T2 (X^{35}) > T4 (X^{51})* ***T3 (X^{214})***	> T1(X^{55}), **T3half (X^{21})**
Group specific	Mandarin	**T3half(X^{21})**	> T1(X^{55})
	Korean	T1(X^{55})	> **T3half(X^{21})**

3.3. Discussion

The results of our perception study showed that, overall, dynamic tones (T2, T3, T4) were perceived as longer than static tones (T1, T3half). This pattern was observed for both groups, indicating that **General Perceptual Bias 1** (dynamic > static) dominates independent of language background, as has also been shown in many previous studies (Faytak & Yu, 2011; Gandour, 1977; Gussenhoven & Zhou, 2013; Šimko et al., 2015; Yu, 2010). Among the contour tones, T2 was perceived as longer than T4, indicating that **General Perceptual Bias 2** (rising > falling) was also in effect, as has been reported elsewhere (Rosen, 1977; Van Dommelen, 1993; Wang et al., 1976). However, differences between language groups were evident, suggesting the importance of linguistic experience. In the following, we consider the source of the observed group differences.

First, between the level tones, the Korean listeners perceived T1 stimuli as longer than T3half stimuli, while the opposite was found for the Mandarin listeners, as suggested by the significant Tone-Group interaction. Consistent with **General Perceptual Bias 3**, according to which syllables with a high-*f0* seem longer than those with a low-*f0*, the results of the Korean group can be taken to support perceptual compensation: due to articulatory constraints, high-*f0* tones are often shorter in production than low-*f0* tones, and this asymmetry is corrected in perception such that high-*f0* tones are overestimated compared with low-*f0* ones, all else being equal. The reversed patterns for the TM group are intriguing. The overestimation of T3half by the TM listeners cannot be attributed to the general perceptual biases, as explained above, nor can it be accounted for by the **Phonetic Bias** hypothesis (6) since in their linguistic experience, T3half is not longer than T1.

One possible explanation comes from its association with T3, which canonically has a complex falling-rising contour. Although the contour of T3 in TM is rarely produced with its full complexity, the canonical forms appear to be clearly stored in the mental representation, exerting a strong influence in processing (Chuang, 2017; Fon et al., 2011; Lu, 2019; Sumner & Samuel, 2005, 2009). These patterns seem to reflect the **Canonicity Bias** (3) which predicts that TM listeners would overestimate T3 due to the complex nature of its tonal representation. Though no significant difference was observed, we did find that T3 stimuli were perceived as marginally longer than T2 ones by the TM listeners. We attribute the longer T3half perception to the lexical association between the members of the T3 category. Although T3half itself is low in pitch and short in duration, the higher-order phonological knowledge of tonal allophones seems to have linked T3half to its unreduced counterpart, T3full, the most complex tone. This seems to have caused the low-*f0* T3half to be ranked above the high-*f0* T1 on the durational scale.

The overall results showed that TM listeners, though partially guided by general perceptual biases, relied heavily on higher-order phonological knowledge (i.e., canonical tonal representation and lexical association of tone variants) in giving durational judgments. Korean listeners, on the other hand, were mainly guided by general perceptual biases and were not biased by their linguistic experience. This leads us to the following questions: Would the Canonicity Bias influence TM speakers' production when imitating tones with various durations? Or would their production simply reflect the temporal

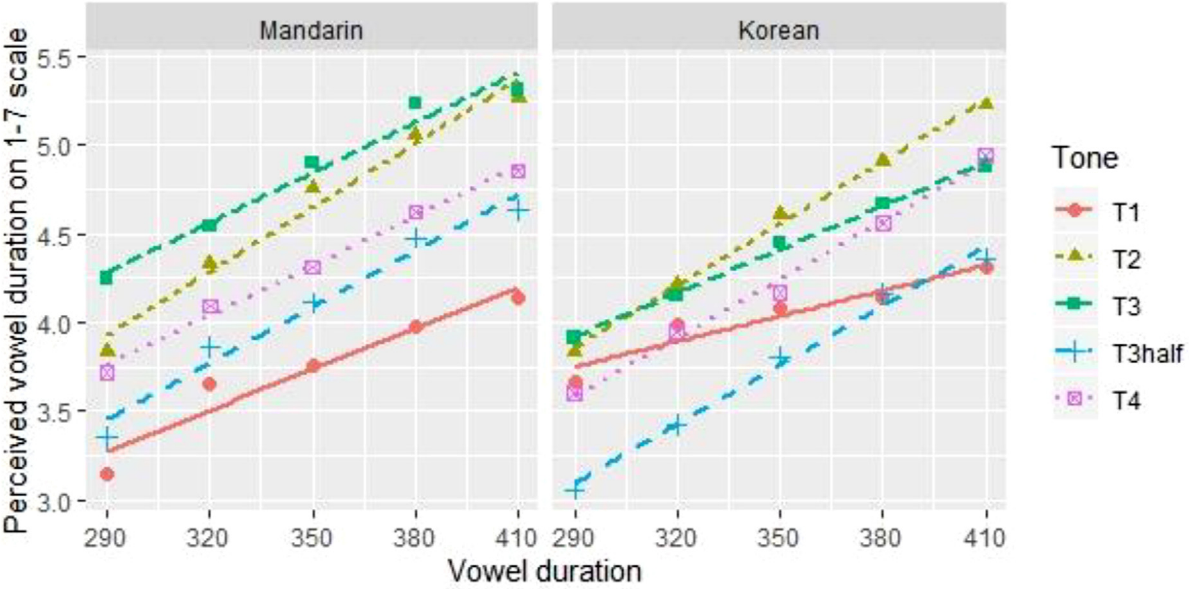

Fig. 6. Fitted regression lines and average perceived vowel duration as a function of Tone and Vowel Duration paneled by Group.

differences of the lexical tones? To answer these questions, we designed an experiment that would probe production.

4. Spontaneous imitation experiment

Following up on the perceived vowel duration experiment in which TM listeners were shown to overestimate the duration of T3 and T3half, an imitation experiment was conducted to test how faithfully TM speakers could imitate stimuli with controlled durations. Previous studies have shown that when imitating auditory prompts, speakers display patterns that can only be explained in terms of their native abstract phonological systems (Kwon, 2019; Mitterer & Ernestus, 2008; Nielsen, 2011). For example, Nielsen (2011) found that, upon being exposed to /p/ with longer VOT, English speakers not only imitated the elongated VOT of the same consonant (e.g., [p]ebble) but produced longer VOT for novel items not previously presented (e.g., [p]illar) as well as for words with different onset consonants (i.e., /k/). Further demonstrating the role of phonology in spontaneous imitation, Kwon (2019) showed that not all phonetic cues are imitated equally. In Korean, aspirated and lenis stops are primarily cued by post-stop *f0,* while stop VOT serves as a secondary cue. When exposed to aspirated stops with elongated VOT (i.e., the secondary cue), Korean speakers not only imitated the longer VOT but also enhanced the post-stop *f0* (i.e., the primary cue). However, VOT was not lengthened when speakers were exposed to an enhanced *f0* cue. These results indicate the influence of higher-order phonological representations during spontaneous phonetic imitation beyond specific phonetic properties in the auditory prompts.

Building on these previous studies, we expect that effects of phonological representations drawn from the native language would be evident in TM speakers' imitations, while the Korean speakers, our control group, would be able to imitate the duration-controlled stimuli more faithfully due to the lack of particular phonological representations of the lexical tones. Specifically, we predict that TM speakers' imitation to be influenced by the implementation of these lexical tones from their experience. Furthermore, their imitations of T3 and T3half are expected to be lengthened to some extent due to the canonical falling-rising contours of the T3 category.

4.1. Methodology

4.1.1. Participants

20 TM speakers (15 female, 5 male; aged 20–23; M = 21) were recruited at National Chiao Tung University to serve as the target group. For the non-tone language baseline group, 21 Korean speakers (9 female, 12 male; aged 20–31; M = 24.5) were recruited at Seoul National University, none of whom had studied Mandarin or spoke Korean dialects with pitch accents. None of the participants in this imitation experiment had participated in the vowel duration judgement experiment. No hearing or speaking deficiencies were reported by any of the participants. All participants were compensated monetarily for their time.

4.1.2. Stimuli

The same 100 stimuli (4 syllables [pa, pi, ta, ti] × 5 tones [T1, T2, T3, T3half, T4] × 5 duration steps [290–320-350–380-410 ms]) that were used in the perceived vowel duration experiment were employed in this task.

4.1.3. Procedure

The experiment was conducted in sound attenuated booths at the universities from which the participants were recruited. The TM participants were recorded using a Marantz PMD661A recording device with an AKG P220 large-diaphragm condenser microphone. The Korean participants were recorded using a Zoom H4 recorder connected to a Shure SM58 microphone. The sampling rate was set as 44.1 kHz and 16 bits for both groups. The procedure followed that of (Davidson, Martin, & Wilson, 2015) with slight modifications. Participants were instructed with verbal and written instructions in their native languages to listen to each stimulus item, played twice with an ISI of 450 ms, and to repeat it as accurately as possible. The 100 stimuli were repeated two times in four blocks using E-Prime (Schneider et al., 2002). The 200 trials were presented in random orders for each participant. They were given 1500 ms before the next trial started. Eight practice trials were presented before the experiment to familiarize participants with the task. The total duration of the experiment was around 12 min. The imitated syllables were then labelled in Praat to obtain their duration. ProsodyPro (Xu, 2013) was employed to obtain *f0* information.

4.2. Results

The data of one Korean participant were excluded from the analysis due to poor recording quality. The aggregated results of all the other participants are shown in Fig. 7. When imitating stimuli of varying durations, less variation across different tones was observed among the Korean speakers compared with the TM speakers. The distinct separations between the lines in the TM speakers' data suggest that phonetic differences in native tonal categories are reflected in TM speakers' production.

To interpret the results, a linear mixed-effects regression model was fitted to the data in R. The dependent variable was the difference in duration between the stimulus and imitated productions. The model included fixed effects for Group (2 levels: TM, Korean), Tone (5 levels: T1, T2, T3, T3half, T4), Duration (5 steps, converted to z-scores to avoid extrapolation beyond the duration steps), and Vowel (/a/= -1 vs. /i/ = 1). The model included random intercepts for Participant as well as by-participant random slopes for Duration.[9] In addition, the model included an interaction term for Group, Tone, and Duration. The statistical model is summarized in (9); for the formulas that generated these results, see Appendix 3.

The significant Duration effect ($\beta = -24.76$, $p < .0001$) without a Duration-Group interaction ($\beta = 4.86$, $p = .22$) suggests that both groups were imitating durational differences in the stimuli, as evidenced by the gradual increase along the x-axis in Fig. 7. However, the imitated tones were generally longer than the input stimuli, and the magnitude of overshoot was not equal across the different tones. Specifically, with T3half and TM as baselines, we found the TM group overshot T1 ($\beta = 34.35$, $p < .0001$), T2 ($\beta = 41.70$, $p < .0001$), and T3

[9] A more complex random structure failed to converge.

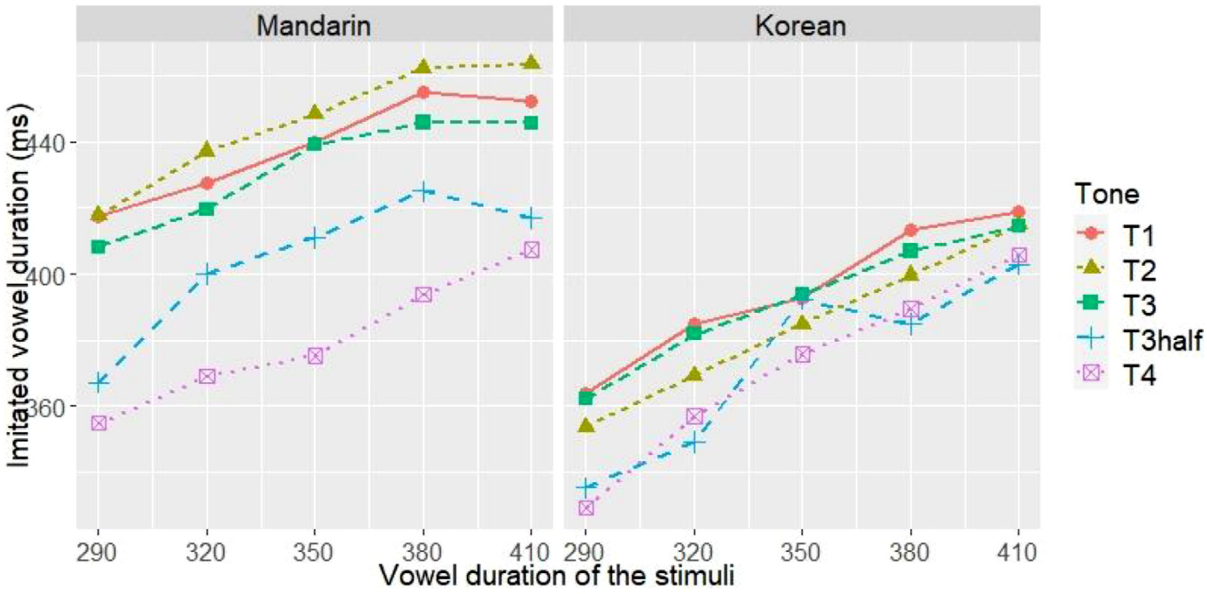

Fig. 7. Average imitated vowel duration as a function of Tone and Vowel Duration paneled by Group.

(β = 27.61, p < .0001) durations more than T3half, while T4 was overshot less than T3half (β = −24.04, p < .0001). Furthermore, the lack of Tone-Duration interactions suggests that the pattern was consistent across different duration steps (T1-Duration: β = −3.85, p = .24; T2-Duration: β = −1.14, p = .73; T3-Duration: β = −3.29, p = .32; T4-Duration: β = 0.72, p = .83).

We also found a marginally significant Group effect (β = −26.08, p = .051), suggesting that the TM group overshot the duration of T3half (M = 54.07 ms) slightly more than the Korean group (M = 22.88 ms). Significant Group-Tone interactions were found for all tones (Korean-T1: β = −12.39, p = 0.008; Korean-T2: β = −29.72, p < .0001; Korean-T4: β = 22.50, p < .0001), except for Group-T3 which was only marginally significant (β = −8.78, p = .061). This suggests that the divergence from the stimuli duration was even larger for the TM speakers' productions of the other tones: M(Korean) = 44.72 ms vs. M(TM) = 88.42 ms for T1; M(Korean) = 34.86 ms vs. M(TM) = 95.77 ms for T2; M(Korean) = 41.79 ms vs. M(TM) = 81.68 ms for T3; M(Korean) = 21.33 ms vs. M(TM) = 30.03 ms for T4. The lack of Group-T3 interaction indicated that the Korean group, similar to the TM group, also imitated T3 as longer than T3half, presumably due to the complexity of dynamic *f0* articulation (also see Section 1.1). The lack of higher order interactions involving Group, Tone and Duration demonstrate that these group differences were robust across different tones and duration steps, as clearly shown in Fig. 7. The longer imitated duration by the TM group is presumably due to the fact that Mandarin tones produced in isolation are generally longer (between 350 ms and 500 ms; see our corpus results in Fig. 2) than the stimuli in this experiment (between 290 ms and 410 ms).

By ordering the tones by degree of overshoot, we see a similar pattern as was found for the TM speakers' production in the corpus study (T1, T2, T3 > T3half), with the exception of T3half and T4. The model showed that the TM speakers overshot T3half to a greater degree compared to T4, contrary to the produced durations in the corpus study in which the duration of T3half was found to be comparable to that of T4.

(9) Summary of fixed effects for imitating vowel duration

	T3half and TM as baselines			
	R^2 = 0.42			
Predictor	B	SE	t	p
(Intercept)	54.06	10.18	5.31	<0.0001
Vowel	4.02	0.74	5.43	<0.0001
Korean	−26.08	13.08	−1.99	0.0510
T1	34.35	3.31	10.38	<0.0001
T2	41.70	3.31	12.60	<0.0001
T3	27.61	3.31	8.34	<0.0001
T4	−24.04	3.31	−7.26	<0.0001
Duration	−24.76	2.82	−8.78	<0.0001
Korean:T1	−12.39	4.69	−2.64	0.0082
Korean:T2	−29.72	4.69	−6.34	<0.0001
Korean:T3	−8.78	4.69	−1.87	0.0610
Korean:T4	22.50	4.68	4.81	<0.0001
Korean:Duration	4.86	3.91	1.24	0.2152
T1:Duration	−3.85	3.31	−1.16	0.2446
T2:Duration	−1.14	3.31	−0.34	0.7309
T3:Duration	−3.29	3.31	−0.99	0.3203
T4:Duration	0.72	3.31	0.22	0.8290
Korean:T1:Duration	−0.70	4.69	−0.15	0.8820
Korean:T2:Duration	−0.87	4.69	−0.19	0.8523
Korean:T3:Duration	−2.53	4.69	−0.54	0.5898
Korean:T4:Duration	1.43	4.68	0.31	0.7596

Several patterns were observed in the imitated *f0* contours. First, the overall *f0* range was lower for the Korean speakers than for the TM speakers. This is presumably due to the fact that there were more male participants in the Korean group. Second, the contours of T3 and T3half were more distinct in Korean speakers' imitation, whereas the two contours were comparable for the TM speakers. Interestingly, when imitating

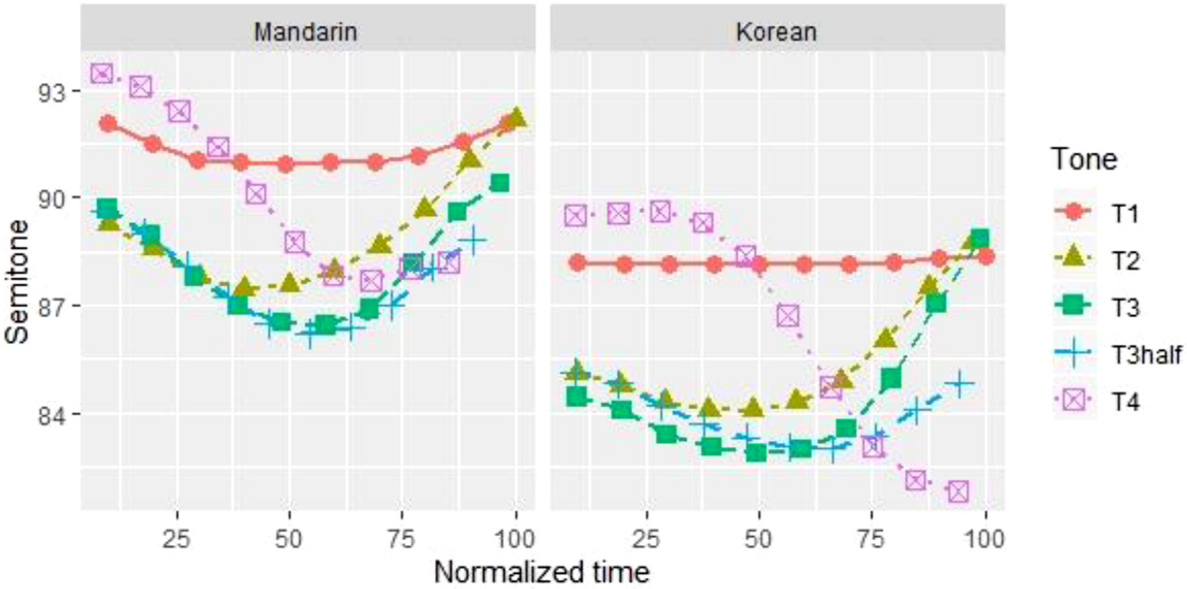

Fig. 8. Time normalized semitones as a function Tone paneled by Group.

T3half, a low-level tone in the auditory stimuli, TM speakers implemented a final-rise, similar to the T3full contour. Attentive readers might have noticed a slight rise in Korean speakers' T3half production as well. This small *f0* dipping found in the Korean data may be attributed to articulatory-phonetic effects (Fig. 8).

4.3. Discussion

The results of the imitation experiment showed a clear effect of language background. When the participants were asked to imitate varying durations of stimuli, the temporal differences across different tones were less variable in the Korean speakers' productions, suggesting that they were more faithful to the phonetic durations of the auditory input. In contrast, TM speakers' produced vowel lengths of the different lexical tones basically followed those in the corpus results, with the exception of T3half and T4: T3half was overshot to a greater degree than T4.

More interestingly, we observed the merging of two *f0* contours, T3full and T3half, in TM speakers' production. This was achieved by slightly shortening T3full and elongating T3half, consistent with the findings in the perceptual experiment. This contrasts with the patterns of the Korean speakers who treated T3half and T3full as separate categories, as indicated by the divergent *f0* contours throughout the vowel. For the TM speakers, overshooting the duration of T3half and adding the final rise to the tonal contour seems to be driven by the association of both T3 tones with a single phonological category.

5. General discussion

This study examined the perceived vowel duration by listeners of Taiwan Mandarin, a language that does not contrast different vowel lengths but does manifest temporal differences in lexical tones. We investigated three hypotheses. First, General Perceptual Biases predicted listeners should perceive vowels with dynamic *f0* as longer than those with static *f0* (T2, T3, T4 > T1, T3half), vowels with rising contours as longer than those with falling contours (T2 > T4), and vowels with high *f0* as longer than those with low *f0* (T1 > T3half). Second, the Phonetic Bias predicted that TM listeners' perceived durations would reflect their linguistic experience; based on the corpus study, lexical tones with longer phonetic durations (e.g., T2) were predicted to be perceived as longer than others (T1 > T3 > T4). Lastly, the Canonicity Bias made a unique prediction in which T3 was likely to be overestimated due to its canonical complex contour, and its variant T3half would likewise be overestimated based on its lexical association to the T3 category.

The data from the Korean listeners and to some extent the data from the TM listeners could be accounted for by the General Perceptual Biases hypothesis. However, it could not explain why TM listeners perceived T3 as the longest, nor could it explain why they perceived T3half, a low-level tone, as longer than T1, a high-level tone. The apparent language-dependent differences in the perception study led us to conclude that perceived duration cannot be entirely explained by a single universal mechanism. The finding that syllables carrying T3 were perceived as the longest by TM listeners, despite their short phonetic duration in production, indicates the significance of the canonical complex dynamicity of the T3 contour. This result provides evidence for the psychological reality of the stored knowledge of abstract canonical representations of lexical tones. The effect of canonicity is further evidenced by TM listeners' overestimation of T3half in production, demonstrating a clear association with the T3 category. These results are indicative of the pervasiveness of canonicity in phonological processing and contribute empirical coverage of a topic which has been limited, to date, to the segmental domain.

The findings of the imitation task further support the influence of canonical representations. The results showed that Korean speakers' imitation patterns were largely governed by the phonetic features of the stimuli (i.e., duration and *f0* contour), while those of TM speakers emerged from a combination of both phonetic and phonological effects. On the one hand, their production largely mirrored the temporal manifestations of lexical tones observed in the corpus study; on the other, T3half was again hyperarticulated, produced as longer than T4, based on its phonological association with T3full. These results echo the findings from previous studies employing similar experimental paradigms that both phonetic and abstract representations are revealed in imitated speech (Kwon, 2019; Mitterer & Ernestus, 2008; Nielsen, 2011; Scarborough et al., 2020).

It is worth noting that in psycholinguistic studies designed to exclude the engagement of higher-order linguistic knowledge, Mandarin listeners do indeed show perceptual patterns similar to those demonstrated by the Korean listeners in the current study, who served as a control group due to their lack of knowledge of a tonal language. In Gussenhoven and Zhou (2013), Mandarin listeners displayed patterns that were in line with predictions based on *perceptual compensation* (Section 1.1). The crucial differences between Gussenhoven and Zhou's study and the current one lie in the properties of the stimuli used in the two experiments. Gussenhoven and Zhou used stimuli produced by a Russian speaker which were manipulated into different *f0* heights/contours which were unfamiliar to Mandarin listeners, while the stimuli in the current study maintained the *f0* heights/contours of naturally produced tokens by a native Mandarin speaker, only manipulated in duration. The stimuli in the current study were presumably similar enough to natural Mandarin speech to have enabled participants to perform a higher level of processing, which in turn drove a strong native-language phonological effect. Gussenhoven and Zhou's study, on the other hand, was designed to tap into primarily phonetic processing and thus induced a pattern consistent with language-independent general perceptual biases.

Based on the empirical foundation established by our corpus study for the temporal differences associated with different lexical tones in TM, the findings of our perception and production experiments suggest that the perception of vowel duration cannot be explained solely by universal mechanisms. Rather, perceived vowel duration is guided by higher-order phonological knowledge from speakers' linguistic experience as well as by general perceptual biases.

CRediT authorship contribution statement

Yu-An Lu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition. **Sang-Im Lee-Kim:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing - review & editing, Funding acquisition.

Acknowledgements

We would like to thank the reviewers of Journal of Phonetics and the attendees at the NCTU 2018 Fall Sound Workshop, LSA 2019, ICPhS 2019, and LabPhon 17 for comments and suggestions on the work. We would also like to thank many research assistants for their help with data collection: Yu-Ming Chang, Pei-Chun Chen, Yangyu Chen, Sarang Jeong, Shao-Jie Jin, Kevin Cheng-Huan Lee, Waan-Rur Lu, Yen-Ju Lu, Yu Nan, Chris Wei-Cheng Weng, and Tzu-Hsuan Yang. This work was supported by MOST106-2410-H-009-031/MOST108-2410-H-009-014 to Yu-An Lu and MOST107-2410-H-009-016-MY3 to Sang-Im Lee-Kim.

Appendix 1. R code for the models reported in (5) for the corpus study

lmer (Produced Duration (z-scored) ~ Tone * Context + (1 + Condition + Tone | Participant) + (1 | Item), data)

Three models with different baselines in Condition and Tone (Isolation/T1, Isolation/T3, Frame/T3) were fitted using the same formula.

Appendix 2 R code and models reported in (7) for the perceived vowel duration experiment

lmer (Perceived Duration (z-scored) ~ Vowel * Duration (centered) + Vowel * Tone + Tone * Duration (centered) * Group + (1 + Duration (centered) + Tone | Participant), data)

Group and Vowel sum coded so that the sum of the weight of each level would be 0: TM = −1 vs. Korean = 1; /a/= −1 vs. /i/= 1.

Two models with different baselines in Tone (T1 and T2) were fitted using the same formula.

Appendix 3 R code and model reported in (9) for the imitation experiment

lmer (Duration Difference (z-scored) ~ Tone * Duration (centered) * Group + Vowel + (1 | Participant), data)

Vowel sum coded so that the sum of the weight of each level would be 0: /a/= −1 vs. /i/= 1.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer (Version 6.0.26). Retrieved from http://www.praat.org.

Bybee, J. (2003). Mechanisms of change in grammaticization: The role of frequency. In B. Joseph & R. Janda (Eds.), *Handbook of historical linguistics* (pp. 602–623). Oxford: Blackwell.

Chang, Y.-L., & Lu, Y.-A. (2016). Onset and tonal effects on perceived vowel duration. *Paper presented at the NACCL 28, Provo, Utah*.

Chao, Y.-R. (1968). *A grammar of spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Chien, Y.-F., Sereno, J. A., & Zhang, J. (2017). What's in a word: observing the contribution of underlying and surface representations. *Language and Speech, 60*(4), 643–657.

Chuang, Y.-Y. (2017). *The effect of phonetic variation on word recognition in Taiwan Mandarin. (Ph. D)*. Taipei, Taiwan: National Taiwan University.

Cumming, R. (2011). The effect of dynamic fundamental frequency on the perception of duration. *Journal of Phonetics, 39*(3), 375–387.

Deng, Dan, Shi, Feng, & Lu, Shinan (2006). The contrast on tone between Putonghua and Taiwan Mandarin. *Shengxue Xuebao (Acta Acoustica), 31*(6), 536–541.

Davidson, Lisa, Martin, Sean, & Wilson, Colin (2015). Stabilizing the production of nonnative consonant clusters with acoustic variability. *The Journal of the Acoustical Society of America, 137*(2), 856–872.

Davis, M. J. (2010). Contrast coding in multiple regression analysis: Strengths, weaknesses, and utility of popular coding structures. *Journal of Data Science, 8*(1), 61–73.

Dawson, C., Aalto, D., Simko, J., & Vainio, M. (2017). The influence of fundamental frequency on perceived duration in spectrally comparable sounds. *PeerJ, 5* e3734.

Duanmu, S. (2007). *The phonology of standard Chinese*. New York: Oxford University Press.

Faytak, M., & Yu, A. (2011). A typological study of the interaction between level tones and duration. *Paper presented at the ICPhS, Hong Kong*.

Fon, J., & Chiang, W.-Y. (1999). What does Chao have to say about tones? A case study of Taiwan Mandarin/赵氏声调系统与声学之联结及量化–以台湾地区国语为例. *Journal of Chinese Linguistics*, 13–37.

Fon, J., Hung, J.-M., Huang, Y.-H., & Hsu, H.-J. (2011). Dialectal variations on syllable-final nasal mergers in Taiwan Mandarin. *Language and Linguistics, 12*(2), 273–311.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language, 84*(3), 474–496.

Gandour, J. (1977). On the interaction between tone and vowel length: Evidence from Thai dialects. *Phonetica, 34*(1), 54–65.

Gussenhoven, C. (2004). Perceived vowel duration. *LOT Occasional Series, 2*, 65–71.

Gussenhoven, C., & Zhou, W. (2013). Revisiting pitch slope and height effects on perceived duration. *Paper presented at the INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Ho, A. T. (1976). The acoustic variation of Mandarin tones. *Phonetica, 33*(5), 353–367.

Jaeger, J. J. (1980). Testing the psychological reality of phonemes. *Language and Speech, 23*(3), 233–253.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience, 13*(8), 1020.

Jun, S.-A. (2005). Korean intonational phonology and prosodic transcription. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 201–229). Oxford: Oxford University Press on Demand.

Kinoshita, K., Behne, D. M., & Arai, T. (2002). Duration and F0 as perceptual cues to Japanese vowel quantity. *Paper presented at the seventh international conference on spoken language processing*.

Köhnlein, B. (2015). The complex durational relationship of contour tones and level tones: Evidence from diachrony. *Diachronica, 32*(2), 231–267.

Kubler, C. C. (1985). The influence of Southern Min on the Mandarin of Taiwan. *Anthropological Linguistics, 27*(2), 156–176.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Test in linear mixed effects model: R package version 2.0-33.

Kwon, H. (2019). The role of native phonology in spontaneous imitation: Evidence from Seoul Korean. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 10*(1).

Kwon, K.-K. (2003). Prosodic change from tone to vowel length in Korean. *Development in prosodic systems*, 67–89.

Lehiste, I. (1976). Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics, 4*(2), 113–117.

Lin, Y.-H. (2007). *The sounds of Chinese*. Cambridge, UK: Cambridge University Press.

Lu, Y.-A. (2019). The effect of dialectal variation on word recognition: A case from Taiwan Southern Min. *Language and Linguistics, 20*(4), 567–600.

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*(1), 168–173.

Myers, J., & Tsay, J. (2003). Investigating the phonetics of Mandarin tone sandhi. *Taiwan Journal of Linguistics, 1*(1), 29–68.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39*(2), 132–142.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. L. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistics structure* (pp. 137–157). Philadelphia: John Benjamins.

Qu, C. (2013). *Representation and acquisition of the tonal system of Mandarin Chinese* (Ph.D). Montreal: McGill University.

Rosen, S. (1977). The effect of fundamental frequency patterns on perceived duration. *Speech Transmission Laboratory—Quarterly Progress and Status Report, 18*, 17–30.

Scarborough, R., Strickler, A., & Nielsen, K. (2020). The effect of linguistic information on f0 imitation. *Paper presented at the LabPhon 17, Vancouver*.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences, 17*(11), 556–564.

Šimko, J., Aalto, D., Lippus, P., W□odarczak, M., & Vainio, M. (2015). Pith, perceived duration and auditory biases: Comparison among languages. *Paper presented at the 18th International Congress of Phonetic Sciences*.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology, 4*, 1015.

Sumner, M., & Samuel, A. G. (2005). Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language, 52*, 322–338.

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language, 60*, 487–501.

Takiguchi, I., Takeyasu, H., & Giriko, M. (2010). Effects of a dynamic F0 on the perceived vowel duration in Japanese. *Paper presented at the Speech Prosody 2010-Fifth International Conference*.

Tseng, S.-C. (2013). Lexical coverage in Taiwan Mandarin conversation. *International Journal of Computational Linguistics & Chinese Language Processing, 18*(1).

Van de Weijer, J., & Sloos, M. (2014). The four tones of Mandarin Chinese: Representation and acquisition. *Linguistics in the Netherlands, 31*(1), 180–191.

Van Dommelen, W. A. (1993). Does dynamic F0 increase perceived duration? New light on an old issue. *Journal of Phonetics, 21*(4), 367–386.

Wang, W. S. Y., Lehiste, I., Chuang, C. K., & Darnovsky, N. (1976). Perception of vowel duration. *The Journal of the Acoustical Society of America, 60*(S1), S92.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49*(1), 25–47.

Wu, F., & Kenstowicz, M. (2015). Duration reflexes of syllable structure in Mandarin. *Lingua, 164*, 87–99.

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

Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis.

Yip, M. (2002). *Tone.* Cambridge University Press.

Yu,A.C.(2010).Tonaleffectsonperceivedvowelduration.*LaboratoryPhonology10,4*(4),151.

Zhang, J. (2000). Phonetic duration effects on contour tone distribution. *Paper presented at the PROCEEDINGS-NELS*.

Zhang, J., & Lai, Y. (2010). Testing the rol of phonetic knowledge in Mandarin tone sandhi. *Phonology, 27*, 153–201.