Phonetics and Phonology in the imitation of tone contrasts

Wei Zhang

Department of Linguistics
McGill University, Montréal

June 2024

A thesis submitted to McGill University in partial fulfillment for the requirements of the degree of Doctor of Philosophy

Abstract

Phonetic imitation of phonetic cues that signal phonological contrasts–e.g. voice onset time, signaling the /t/-/d/ contrast–is generally non-linear. However, previous work on how phonological categories mediate imitation has focused more on cues to segmental contrasts (i.e., vowels and consonants), neglecting suprasegmental contrasts (e.g. tones). This dissertation examines phonological effects and phonetic variations in imitation, for two cues to tonal contrasts: F0 and duration.

This dissertation contains three studies. The first study examines imitation of the flat-falling tone contrast in Mandarin by native speakers. Two phonetic cues to the contrast, F0 fall range (the primary cue) and duration (a weak cue), are manipulated to create a set of flat-falling continua. The distribution of imitations is analyzed by comparing the fit of Gaussian models containing one component (i.e., linear model) and two components (i.e., categorical model). The two-component model fits the F0 imitation data better, suggesting more categorical imitation. However, the one-component model fits the duration imitation data better, suggesting a more linear pattern. These results indicate that F0 imitation is subject to phonological mediation, while duration imitation is not.

The second study further explores duration imitation of tones of Taiwanese Southern Min. Previous work has mostly found that duration imitation is not mediated by phonological contrasts, but using cases where duration served as a weak cue. It is unclear if duration imitation remains linear when it is an important cue, as is thought to be the case for Taiwanese Southern Min. A first experiment finds that duration is important for distinguishing T3 (mid-register checked tone, with a final glottal stop) from T33 (mid-register unchecked tone). A

second experiment examines imitation of a T3-T33 continuum created by varying the duration. The imitation data is best modeled by a one-component model, suggesting that it is not mediated by the T3-T33 contrast, despite acting as a primary cue to it.

The third study compares imitation of the flat-falling contrast between native Mandarin speakers and native English speakers without tone experience. We explore if speakers lacking lexical tone phonology imitate F0 range linearly or categorically, to understand the type of category which can mediate imitation–if English speakers show categorical behavior, it cannot be due to tonal categories. Results show that both groups imitate the continuum as a mix of flat and falling categories while also tracking some within-category F0 variability, indicating that both pre-existing categories and within-category variation are sources of imitation. However, Mandarin speakers' imitation is more categorical and shows hyper-articulation, suggesting it is mediated by linguistic contrasts, while the English group's imitation seems to be mediated by psychophysical contrasts.

In summary, this dissertation explores phonetic imitation of f0 and duration as cues to lexical tones. The three studies show that F0 imitation can be mediated by both phonological and psychophysical contrasts, whereas duration imitation is not mediated by phonological contrasts. These results reveal that phonological mediation in imitation is dimension-dependent, shedding light on the mechanism of phonetic imitation, ultimately enriching our knowledge of language processing and production mechanisms.

Resumé

L'imitation phonétique des indices phonétiques signalant les contrastes phonologiques - par exemple, le délai d'établissement du voisement, signalant le contraste /t/-/d/ - est généralement non-linéaire. Cependant, les travaux précédents sur la manière dont les catégories phonologiques médiatisent l'imitation se sont principalement concentrés sur les indices des contrastes segmentaux (c'est-à-dire les voyelles et les consonnes), en négligeant les contrastes suprasegmentaux (par exemple, les tons). Cette thèse examine les effets phonologiques et les variations phonétiques dans l'imitation, en se concentrant sur deux indices de contrastes tonals: F0 et la durée.

Cette thèse comprend trois études. La première étude examine l'imitation du contraste tonal "plat-descendant" en mandarin par des locuteurs natifs. Deux caractéristiques phonétiques signalant le contraste, la plage de chute de F0 (l'indice principal) et la durée (un indice faible), sont manipulées pour créer un ensemble de continus "plat-descendant". La distribution des imitations est analysée en comparant l'ajustement de modèles gaussiens contenant une composante (modèle linéaire) et deux composantes (modèle catégoriel). Le modèle à deux composantes s'ajuste mieux aux données d'imitation de F0, suggérant une imitation plus catégorielle. Cependant, le modèle à une composante s'ajuste mieux aux données d'imitation de la durée, suggérant un schéma plus linéaire. Ces résultats indiquent que l'imitation de F0 fait objet d'une médiation phonologique, tandis que l'imitation de la durée n'en fait pas.

La deuxième étude explore davantage l'imitation de la durée des tons du Minnan du Sud du Taïwan. Les travaux précédents ont majoritairement conclu que l'imitation de la durée

n'est pas médiatisée par les contrastes phonologiques, en se basant sur des cas où la durée servait d'indice faible. Il n'est pas clair si l'imitation de la durée reste linéaire lorsqu'elle est un indice important, comme cela est supposé pour le Minnan du Sud du Taïwan. Une première expérience montre que la durée est importante pour distinguer T3 (ton de registre moyen avec une occlusive finale) de T33 (ton de registre moyen sans occlusive finale). Une deuxième expérience examine l'imitation d'un continuum T3-T33 créé en faisant varier la durée. Les données d'imitation sont mieux modélisées par un modèle à une composante, suggérant qu'elles ne sont pas médiatisées par le contraste T3-T33, malgré son rôle d'indice principal.

La troisième étude compare l'imitation du contraste "plat-descendant" entre des locuteurs natifs du mandarin et des locuteurs anglophones sans expérience tonale. Nous explorons si les locuteurs sans phonologie tonale imitent la gamme de F0 de manière linéaire ou catégorielle, afin de comprendre le type de catégorie qui peut médiatiser l'imitation. Les résultats montrent que les deux groupes imitent le continuum comme un mélange de catégories "plat" et "descendant", tout en suivant également une certaine variabilité de F0 à l'intérieur des catégories, indiquant que les catégories préexistantes et la variation intra-catégorielle sont sources d'imitation. Cependant, l'imitation des locuteurs mandarins est plus catégorielle et montre une hyper-articulation, suggérant qu'elle est médiatisée par des contrastes linguistiques, tandis que celle du groupe anglais semble être médiatisée par des contrastes psychophysiques.

En résumé, cette thèse explore l'imitation phonétique de F0 et de la durée en tant qu'indices des tons lexicaux. Les trois études montrent que l'imitation de F0 peut être médiatisée à la fois par des contrastes phonologiques et psychophysiques, tandis que l'imitation de la durée n'est pas médiatisée par les contrastes phonologiques. Ces résultats

révèlent que la médiation phonologique dans l'imitation dépend des caractéristiques, éclairant ainsi le mécanisme de l'imitation phonétique et enrichissant nos connaissances sur les mécanismes de traitement et de production du langage.

Acknowledgement

This dissertation would not have been possible without the support of many people. First, I would like to express my gratitude to Meghan Clayards, who has been my advisor since my study at McGill. Meghan has inspired my research with her profound knowledge and high-level academic standards. She has also been incredibly supportive and responsible, showing patience and care for both my research and personal life. I feel very fortunate to be her student.

I'm also very grateful to my advisor, Morgan Sonderegger, and my co-supervisors, Michael Wagner and Francisco Torreira. Morgan's continuous and generous guidance has supported me throughout my research and writing, both within and beyond this dissertation. Specifically, this dissertation would not have been possible without Morgan, as all the statistical modeling in it is guided by Morgan. I am also thankful to Michael for his guidance of my first project at McGill. I am impressed by his vast ranges of knowledge. He has been a role model for me and inspires me to pursue lifelong learning. I have enjoyed all the discussions with Francisco, who offered valuable comments on Chapter 2 of this dissertation and provided significant encouragement during my first year at McGill.

I'm further grateful to my committee member, Yu-An Lu, for her essential input and funding support on Chapter 3, as well as for all the pleasant conversations we have had.

I would like to thank all our members of the Department of Linguistics. Special thanks to the instructors who equipped me with valuable knowledge in linguistics: Heather Goad, Bernhard Schwarz, Jessica Coon, Martina Martinović, Marcelo Vieira. I'm also grateful to my cohorts and colleagues in Morgan's, Meghan's and Michael's labs: Connie Ting,

Jonathan Palucci, Beini Wang, Xuanda Chen, Alex Zhai, Irene Smith, Massimo Lipari, Claire Honda, Amanda Doucette, Jeanne Brown, Terrance Gatchalian, Alvaro Zurita, Jing Ji, Bing'er Jiang. I truly enjoy the time spent with all of you, which made McGill an enjoyable place to be a student. Additionally, I'm thankful to my best friends in China, Ning, Yuning and Rong, with whom I talked about anything, anytime, every day.

Finally, I could not have completed this dissertation without the support of my family members. I am deeply grateful to my parents for their unconditional love and support. A special thank you to Tao for his love, patience, support, and encouragement.

Table of Contents

List of Figures

List of Tables

Contribution of authors

The studies presented (in chapter 2-4) in this thesis have been prepared for publication in peer reviewed journals.

Chapter 2 has been published in the *Journal of Phonetics* and is co-authored with Meghan Clayards and Francisco Torreira. Wei Zhang conceived this study under the guidance of Meghan Clayards and Francisco Torreira. Wei Zhang collected the data and then analyzed it with the guidance of Meghan Clayards, Morgan Sonderegger, and Francisco Torreira. Wei Zhang completed the original draft of the manuscript, with Meghan Clayards and Francisco Torreira providing comments for revision.

Chapter 3 has been prepared for submission and is co-authored with Yu-An Lu, Morgan Sonderegger, and Meghan Clayards. The study was conceived by Wei Zhang and Yu-An Lu under the guidance of Meghan Clayards. Wei Zhang collected the data and then analyzed it with the guidance of Meghan Clayards and Morgan Sonderegger. Wei Zhang drafted the manuscript, which was then revised according to comments from Yu-An Lu, Meghan Clayards, and Morgan Sonderegger.

Chapter 4 has been submitted to a journal and is co-authored with Meghan Clayards and Morgan Sonderegger. The study was conceived by Wei Zhang and Meghan Clayards. Wei Zhang collected the data and analyzed it with the guidance of Morgan Sonderegger and Meghan Clayards. Wei Zhang drafted the manuscript, which was then revised by Wei Zhang, Meghan Clayards, and Morgan Sonderegger.

Chapter 1

General Introduction

1.1 Imitation in a broad sense

Imitation, briefly defined as the copying of behavior, is of interest in many fields such as psychology, anthropology, biology and linguistics. Imitation is a very basic capacity and an inborn ability of human beings and even animals (Meltzoff & Moore, 1997; Custance et al., 1995). The ability of imitation is present even in the first hours after birth – the imitation of adults' facial expressions (such as sticking out the tongue, protruding the lips) was found for infants as young as 60 minutes old (Meltzoff & Moore, 1977). Imitation can be seen as the way by which newborns map 'what they see' onto 'their own body movement', thus it has been claimed as evidence of an intrinsic link between perception and production of an act, and the innate skill of connecting 'self' and 'others' (Meltzoff, 2011). Imitation, albeit covert, has also been proposed to be an underlying process to facilitate action perception, action prediction and action execution (Pickering & Garrod 2013).

Vocal imitation has been found for infants of 12 to 20 weeks (Kuhl & Meltzoff, 1982, 1996). In Kuhl & Meltzoff (1996), the infants' vocal imitation (or babbling) of the /a/, /i/, /u/ stimuli were recognizable to phonetically trained listeners. Vocal imitation is not only crucial for first language acquisition (Kuhl & Meltzoff, 1982, 1996), but also very useful for children and adults in their language learning (Holley & King, 1971; Jensen & Vinther, 2003),

vocal music training (Brophy, 2001; Pfordresher et al., 2022), acting or joking expertise (Zetterholm, 2002, 2007), and language comprehension (Pickering & Garrod 2013; Adank et al., 2013). For example, it has been used in computer-assisted prosody training for L2 learners (De Meo et al., 2013). Imitation of sentences in an unfamiliar accent facilitated comprehension of subsequent sentences in that accent in imperfect listening conditions such as the context of noise (Adank et al., 2013).

In addition to intentional imitation, unintentional imitation behavior occurs in many cases as well. For example, people sometimes spontaneously imitate the facial expressions, gestures or certain mannerisms of others after observing these actions (Dimberg et al., 2000). This phenomenon was named the "chameleon effect" in psychology (Chartrand & Bargh, 1999). Neuroscientists have agreed that the mirror neuron system is the neural basis of imitation behavior. Empirical evidence has been found that when people are observing an action the mirror neurons for the execution of the same action are activated (Rizzolatti et al. 1996; Watkins et al., 2003; Rizzolatti et al., 2001; Iacoboni et al., 1999; Wilson et al., 2004; Heyes, 2001 and others), which facilitates the tendency of imitating the action (Wolpert et al., 2001; Bien et al., 2009). For example, Watkins et al. (2003) revealed that both the visual (viewing speech-related lip movements) and the auditory (listening to speech) perception of speech can result in higher excitability of the motor units involved in speech production (e.g., those for controlling the lip muscles).

In the field of linguistics, it has been found that interlocutors in a dialogue tend to automatically align their syntactic and semantic representations (Pickering & Garrod, 2004). Furthermore, studies have shown that adults unconsciously converged to a different regiolect (i.e., a variety of a standard language in a different region) after a short period of exposure

(Delvaux & Soquet, 2007; Sancier & Fowler, 1997). Therefore, unintentional imitation, which is also named convergence or accommodation, is considered to play an important role in the transmission and propagation of language change through speech communities (Niedzielski & Giles, 1996; Fitch 2000; Lin et al., 2021).

1.2 Phonetic imitation and phonological mediations

This dissertation focuses on imitation at the phonetic level. Phonetic imitation has been extensively studied in the past few decades (Babel, 2009; Chistovich et al., 1966; Flege & Eefting, 1988; Fowler et al., 2003; Goldinger, 1998; Kent, 1974; Mitterer & Ernrstus. 2008; Kim & Clayards, 2019; Kwon, 2019; MacLeod & Di Lonardo Burr, 2022; Lin et a;., 2021; Nielsen, 2011; Nielsen, 2014; Yu et al., 2013; Sato et al., 2013; Dilley, 2010; Pardo, 2006; Pardo et al., 2017; Pierrehumbert & Steele, 1989; Zellou et al., 2016; Zellou et al., 2017; Schertz & Johnson, 2022; Paquette-Smith et al., 2022; Schertz & Paquette-Smith, 2023). Among these phonetic imitation studies, VOT, speech rate (or duration), F0 (including height and contour) and vowel formants were the commonly examined variables. The stretch of speech over which imitation is investigated can be a syllable, a word, a phrase, a sentence, or a conversation. The method of eliciting the imitation encompasses shadowing, where imitation is implicit or subconscious to the participant, and forced imitation, which is explicit to the participant.

Previous literature has discussed the factors that affect the tendency and direction of phonetic imitation (see Kovaříková 2023 for a review). For example, individual-level traits. In a study by Yu et al. (2013), the extent of VOT imitation in a shadowing task was found to

be modulated by individual differences in personality traits (specifically, openness) and cognitive traits related to attention switching. In another study, Lewandowski and Jilka (2019) found that phonetically talented learners converged more towards the native English speakers in a conversational accommodation task than the less talented learners. Second, social and communicative factors have been examined. According to the Communication Accommodation Theory (CAT), imitation/convergence is a means of minimizing social distance from the interlocutor, or demonstrating a desire to belong to a particular social group (Giles, 1973; Coupland et al., 1991). The impact of subjective attitudes toward the talker (Lin et al., 2021; Yu et al., 2013) and the gender of both the speaker and imitator on phonetic imitation has also been extensively discussed (see Pardo et al., 2017 for a review). Furthermore, age effects have been examined. Nielsen (2014) compared the extent of VOT imitation among three age groups: preschoolers, third-graders and adults, and found that children imitated the VOTs to a greater degree compared to adults. Such findings are consistent with the common observation that young children are more likely to obtain a native-like accent when exposed to an L2 environment than adult. However, other studies failed to observe more imitation in children than adults (Paquette-Smith et al., 2022, Schertz & Johnson, 2022).

Another group of studies investigated linguistic factors in phonetic imitation, namely the mediation effects of phonological representations, which is the main focus of interest in this thesis. Phonetic imitation includes a perception phase and a production phase, and differences in imitation can be rooted in either or both of these phases (Vallabha and Tuller, 2004). Early studies have suggested that the perception of a phonological contrast can be non-linear between two categories when the critical cue changes linearly (i.e., categorical

perception), at least for some tasks (Liberman et al., 1957). Thus, it is intuitive to suggest that phonetic imitation may also exhibit non-linearity. Specifically, when a critical phonetic cue changes linearly between two categories, phonetic imitation may not perfectly track the linear steps and instead exhibit clustering due to the non-linear perception phase mediated by the phonological contrast. That is to say, how precisely a cue value is imitated is partially dependent on its distance to the categorical boundary or the distance to the prototypes. This intuition has been supported by many studies (Chistovich et al., 1966; Flege & Eefting, 1988; Kent, 1973, 1974). Among others, Flege and Eefting (1988) investigated the VOT imitation for three groups of adult participants: monolingual Spanish speakers, monolingual English speakers and late childhood Spanish-English bilinguals. They used a /da/ to /ta/ continuum, which was created by varying VOT from -60 ms to 90 ms with the step size of 10 ms. Results showed that all three groups of participants imitated the steps of the continuum in a categorical way – both monolingual groups imitated the VOTs into two modes/clusters and the stimuli between the modes were imitated with larger VOT variations which suggest less precise imitations. In particular, the two modes of the Spanish monolinguals were -88 ms and 18 ms while those of the English monolinguals were 22 ms and 70 ms. The differing modes for monolingual Spanish and English speakers well reflected the effect of native phonology: phonemes /d/ and /t/ are realized as pre-voiced [d] (corresponding to the mode of -88 ms) and unvoiced [t] (corresponding to the mode of 18 ms) respectively in Spanish, whereas they were realized as devoiced [d] (corresponding to the mode of 22 ms) and aspirated [t^h] (corresponding to the mode of 70 ms) respectively in English. Furthermore, the bilingual group imitated the VOTs into all three modes (-82 ms, 26 ms and 70 ms), suggesting that both the L1 and L2 phonologies can mediate phonetic imitation.

In addition, Nielsen (2011) manipulated the VOT of the English /p/ into three levels (normal, lengthened, and shortened) and found that native English participants consistently imitated the lengthened VOT of /p/ after being exposed to stimuli with lengthened VOTs, however, they did not imitate the shortened VOT. This observation suggests that even within the same phonological category, the tendency to imitate extended and reduced cues may differ. The inhibited imitation of the shortened VOT could be due to the linguistic effect of the /p-b/ contrast, as the shortened VOT makes the distinction between the voiced and voiceless stops more ambiguous, and speakers may avoid producing ambiguous phonemes. Alternatively, as discussed by Nielsen (2011) and explored more recently by Schertz and Paquette-Smith (2023), who found evidence that participants could imitate shortened VOT, albeit to a lesser degree than the lengthened ones, the asymmetry in the imitation of extended and shortened VOT could also be due to differences in their perceptual/auditory salience. Specifically, shortened VOT may be perceptually less salient than extended VOT, resulting in reduced/limited imitation of the less salient cue. Subsequent studies also discussed the impact of perceptual salience on imitation performance (Podlipskỳ & Simácková, 2015; Nielsen & Scarborough, 2015, Kim & Clayards, 2019; Schertz & Johnson, 2022; Schertz and Paquette-Smith, 2023). Among others, Podlipskỳ & Simácková (2015) even argued that perceptual salience was a very strong predictor of imitative performance. However, since perceptual salience is a very complex concept and it is notoriously difficult to measure (see MacLeod, 2015 for a review), the extent and mechanisms by which it influences phonetic imitation remain open questions.

In summary, phonetic imitation can be influenced by individual-level differences, social factors, phonological mediation (i.e., linguistic effects), and possibly perceptual salience. In

particular, a number of studies have confirmed the mediation of phonological contrast on phonetic imitation, and this mediation effect is also the most relevant factor for this dissertation.

1.3 Supra-segmental superiority in phonetic imitation

In addition to the above-mentioned factors, a general claim about phonetic imitation is that supra-segmental dimensions are easier to imitate than segmental dimensions (Garnier et al., 2013; Pardo, 2010; Sato et al., 2013). Earlier studies examining the social factors of spontaneous imitation and the imitation of dialect or speaker characteristics have typically used longer stretches of speech, such as sentences or conversations. These studies have generally found that F0, intensity, and speech rate tend to converge to that of the speaker (D'Imperio & Sneed, 2015; Giles et al., 1991; Pardo, 2010; Zetterholm, 2007), whereas the convergence of formants has been less consistent (Babel, 2009; Pardo, 2010). These findings led subsequent studies to hypothesize that supra-segmental dimensions are easier to imitate than segmental dimensions (Garnier et al., 2013). For example, Kim and Clayards (2019) carried out imitation experiments using a vowel continuum which varied along two dimensions: vowel quality (the first and second formants) and vowel duration. Results showed that vowel duration imitation was fairly linear – the duration difference between consecutive steps was well reflected in the imitation without any merging or variation discontinuities induced by the vowel categories. In contrast, the imitation of the vowel quality was non-linear – there were more variations for the intermediately ambiguous steps than the prototypical steps. In addition, some researchers have observed an apparent asymmetry

between the phonetic imitation of F0 and F1. Sato et al. (2013) investigated the imitation of F0 and F1 in three English vowels /i, e, ɛ/. Results revealed that F0 was more accurately imitated than F1 in both spontaneous and explicit imitation tasks. This comparative analysis of the imitation performance of the two acoustic dimensions provides compelling evidence in support of the hypothesis that supra-segmental dimensions are generally easier to imitate than segmental dimensions.

However, it is not the case that 'supra-segmental dimensions' are always easily imitated. Pierrehumbert and Steele (1989) manipulated the F0 peak position of a rise-fall-rise intonation pattern in English – a suprasegmental phenomenon – in an imitation experiment. Results showed that participants imitated early and late peaks well, but were not able to accurately imitate the intermediate peaks. They concluded that there was a boundary within the F0 peak continuum, and that the timing of the F0 peak was interpreted as a binary distinction (possibly corresponding to two categories: L*+H and L+H*) rather than a continuous variable. This study showed that the imitation of F0 peak position, which could be seen as a 'supra-segmental dimensions' in a broad sense, was mediated by the proposed L*+H vs. L+H* distinction, further calling into question the 'supra-segmental dimensions' superiority hypothesis.

Additionally, the definition of 'supra-segmental dimensions' can be vague in some cases, which can make the general hypothesis less clear. Phonologically, supra-segmental phenomena and segmental phenomena are distinguishable (Goldsmith, 1990). The segmental phenomenon concerns the contrastively phonemic categories such as vowels and consonants. As the name suggests, a suprasegmental phenomenon often involves stretches of speech that span multiple segments (Beckman, 1996), for example, the domain of a prosodic phrase and

the presence or absence of an accent. In general, F0, speech rate (duration) and intensity are the common means of marking suprasegmental phenomena. Due to the large body of acoustic studies of suprasegmental phenomena, these dimensions thus tend to be called 'suprasegmental dimensions' in many studies (Cutler et al., 1997). This is basically where the concept 'suprasegmental dimensions' originate from.

Nevertheless, there are actually cases where the so-called 'suprasegmental dimensions' serve as cues for segmental contrasts, or conversely, where 'segmental dimensions' are cuing suprasegmental contrasts, making the concepts of suprasegmental or segmental dimensions less clear cut. For example, the 'suprasegmental' F0 of the following vowel onset could serve as a cue for the voicing contrast of the preceding stop, in addition to VOT (Whalen et al., 1993). Likewise, duration could be a main cue for long and short vowel contrast in languages such as Swedish (Elert, 1964). In addition, segments differ intrinsically in 'suprasegmental dimensions' such as F0, duration and intensity (Hillenbrand et al., 1995; Lehiste & Peterson, 1959) – low vowels in general have lower intrinsic F0, longer intrinsic duration and higher intrinsic intensity than high vowels. Since they are cuing segmental contrasts, F0, duration and intensity in these cases are arguably segmental dimensions.

An example of a 'supra-segmental dimension' playing a segmental role thus being an arguably 'segmental dimension' is the English /ɛ/-/æ/ continuum (from 'head' to 'had') tested in Kim and Clayards (2019), discussed above. They found that for formants, the imitation of the ambiguous steps was more variable than the prototypical steps, however, the imitation of vowel duration did not show this nonlinearity. The results seemed to be compatible with both the linguistic effect (in that the important cue - formants - were nonlinearly imitated) and the easy-to-imitate hypothesis for suprasegmental dimensions (in

that the suprasegmental cue - duration - was linearly imitated) if vowel duration was regarded as a suprasegmental cue here. Therefore, it is worth rethinking what drives the distinction in the imitation of the two cues in Kim and Clayards (2019).

1.4 Overview

This dissertation focuses on the phonological mediation effect in phonetic imitation, which previous literature has shown to be complex and not fully understood. Specifically, phonological mediation is clear in the imitation of some cues and contrasts (e.g., VOT for/t/-/d/ in Flege and Eefting (1988)) but not for others (e.g., vowel duration for /ɛ/-/æ/ in Kim and Clayards (2019)). This dissertation examines the imitation of duration and F0 as cues for lexical tones in Mandarin and Taiwanese Southern Min, which have been relatively less investigated and yield mixed results, to further understand phonological mediation in imitation. Moreover, Wang (1976) proposed that the perception of Mandarin flat-falling tonal contrast is guided by both linguistic categories, expected in native speakers, and psychophysical categories, relevant to all speakers. However, it remains unclear whether the F0 imitation is mediated by linguistic categories, psychophysical categories, or both. This dissertation also aims to explore which levels of categories can mediate imitation to enhance our understanding of the mechanisms of phonetic imitation.

Furthermore, we expect to probe to what extent the representation of a category (or the category structure) is reflected in imitation through the phonological mediation effect. Over the past decades, extensive studies have explored the process of speech perception, specifically how human listeners convert the (continuous) speech signal into discrete categories. As

reviewed in McMurray (2022), it's generally agreed that speech perception (or speech recognition) is a categorization activity. Notably, increasing evidence shows that speech perception retains much more information beyond just the symbolic category, but also within-category details. The fine-grained perceptual nature reveals that representation of a phonological category includes rich and multifaceted information rather than a mere linguistic symbol. For example, the "perceptual magnet theory" suggests that instances within a phonological category are not perceptually equivalent, with some being prototype and others not (Kuhl, 1991). Such graded instances included in a phonological category can be shown in perceptual results from goodness rating tasks (Kuhl, 1991; Miller and Volaitis, 1989; Strange et al., 2004), Visual Word Paradigm experiments (McMurray et al., 2008; McMurray et al., 2002), and Event Related Potential studies (Kapnoula and McMurray, 2021; Toscano et al., 2010). Nevertheless, the extent to which phonetic imitation can reflect the representation of a category, and how that links to our understanding of speech perception, has been less studied in the past. Three studies are carried out in this dissertation to explore the following two research questions.

First, this dissertation examines two possible mechanisms to account for the linear imitation of duration observed in Kim and Clayards (2019). One possibility is related to the general hypothesis – duration is innately supra-segmental even when it is cuing a vowel contrast. Since vowel quality is segmental, duration should be linearly imitated and vowel quality should be non-linearly imitated. The other possibility is about the phonological mediation effect. Given that duration is a less important cue to the examined vowel contrast compared to vowel quality, the mediation on duration is weak. As a result, only the primary cue, vowel quality, is observed to be subjected to phonological mediation, whereas the less

important cue, duration, is imitated linearly. This question was investigated in Chapter 2 and Chapter 3.

To tease apart the two possibilities, Chapter 2 examined the imitation of the flat vs. falling tonal contrast in Mandarin. As will be introduced in more details in Section 2.1.4, the flat tone and the falling tone in Mandarin are primarily distinguished by F0, with duration as a less important cue (Chao, 1965; Blicher et al., 1990). In this case, we controlled for both cues to be supra-segmental, as they are cuing lexical tone, an uncontroversially supra-segmental phenomena. Therefore, if the innate supra-segmental quality of duration (the first possibility mentioned above) is the reason for the linear imitation in Kim and Clayards (2019), we should observe that both F0 and duration are linearly imitated since they are both supra-segmental in nature. If, on the contrary, as in Kim and Clayards (2019) the duration is linearly imitated because it is a less important cue to the vowel contrast, then we should observe similar results to those of Kim and Clayards (2019): non-linear imitation of the primary cue F0 and linear imitation of the less important cue duration.

Chapter 3 further investigated the two possibilities, through the imitation of the mid-checked vs. mid-unchecked tonal contract in Taiwanese Southern Min (TSM). As will be reported in Section 3.2, in contrast to Chapter 2 and Kim and Clayards (2019) where duration serves as a weaker cue to the investigated phonological contrast, duration plays a very important role for the contrast of TSM mid-checked vs. mid-unchecked tones. In this way, the phonological effect and the supra-segmental effect are more directly compared on the dimension of duration. If the innate supra-segmental quality of duration is the reason for the linear imitation in Kim and Clayards (2019), we should observe that duration is linearly imitated across the TSM tonal contrast. If, instead, in Kim and Clayards (2019) the duration

is linearly imitated because it is a less important cue to the vowel contrast, then we should observe a non-linear imitation of duration across this TSM tonal contrast.

Second, this dissertation explores which levels of category can mediate imitation by examining the effect of native phonology. Native phonology shapes the imitation distribution–as introduced in Section 1.2, Flege and Eefting (1988) compared the imitation of VOT of /ba/-/pa/ among Spanish monolinguals, English monolinguals and English-Spanish bilinguals, and found that the way in which the VOT values were distorted in imitation was restricted to their obtained phonology (i.e., the VOT values were imitated into two modes of -88 ms and 18 ms for Spanish monolinguals, 22 ms and 70 ms for English monolinguals, and into three modes for English-Spanish bilinguals). Chapter 4 compares the imitation distributions of the flat-falling tonal continuum in Mandarin by native and non-native speakers of Mandarin. Non-native speakers without any experience of lexical tone, for example, English monolinguals, do not possess the flat-falling phonology for lexical tones. However, as human speakers, they have the psychophysical or prosodic distinction of flat vs. falling F0 contours. If they imitate the steps in the flat-falling continuum in a linear way, it indicates that that phonological categories exclusively mediate phonetic imitation. Alternatively, if they imitate the steps in the continuum in a non-linear way, it indicates that that phonological categories do not exclusively mediate phonetic imitation. Instead, imitation is also mediated by psychophysical categories (or prosodic categories, although Section 4.1 discusses reasons why prosodic categories are a less likely explanation).

Methodologically, the linearity and categoricalness of the imitation data are analyzed by comparing the fit of two cognitive models:

- The linear regression model, which assumes speakers linearly track phonetic cues.
- The categorical regression model, which assumes imitation reflects underlying categories.

The method builds on the approach taken by Massaro & Cohen (1983) in modeling rating responses on a continuous linear scale for stimuli from continua ranging between two categories. Specifically, we further compare the contribution of each model in explaining variations in the imitation distribution by model averaging (details provided in Section 4.4), and interpret the weight of each model as the degree of the linearity and categoricalness in the imitation.

Chapter 2

Study 1: Native imitation of Mandarin flat-falling tonal continua

This chapter is a published journal paper: Zhang, W., Clayards, M., & Torreira, F. (2023). Phonological mediation effects in imitation of the Mandarin flat-falling tonal continua. *Journal of Phonetics*, 101, 101277. https://doi.org/10.1016/j.wocn.2023.101277

Abstract

Phonetic imitation has been found to be mediated by phonological contrast. For features whose values vary around a phonological prototype, the imitation is distorted by the phonological category, i.e., the imitation is nonlinear. This phonological mediation effect was mostly found in segmental features such as VOT and formants. Supra-segmental features, on the contrary, are generally found to be easy to imitate, i.e., the imitation is linear. Nevertheless, whether the phonological effect exists in the imitation of supra-segmental features is not fully understood. This study, through an imitation experiment of Mandarin Flat-falling tonal continua, examined whether a supra-segmental feature would be linearly imitated when it is the primary cue (F0 range) and the non-primary cue (duration) to the tonal contrast, respectively. Results showed that F0 range imitation was non-linear while duration imitation was linear. This reveals that the phonological effect is stronger in mediating imitation than would be predicted by the general hypothesis that supra-segmental features are easier to imitate.

2.1 Introduction

Phonetic imitation involves perception, memorization and production processes, so it can provide a window into how information in the speech signal is processed by the subject. Therefore, it is also of interest which factors affect imitation performance. On the one hand, it has been argued that individuals tend to be better at imitating supra-segmental or prosodic features [1](e.g., F0, duration) than features cueing segmental contrasts (e.g., VOT, formants) (Pardo, 2010; Sato et al., 2013, Garnier et al., 2013). On the other hand, phonological categories may be another factor that limits how well a feature will be imitated (Flege and Eefting, 1988; Kim & Clayards, 2019; Nielsen, 2011). Specifically, speakers tend not to produce ambiguous tokens near a category boundary. These two factors seem to make conflicting predictions for imitation performance in the case where a feature is supra-segmental and is an important cue to a phonological contrast. The current study aimed to test these two predictions using a Mandarin lexical tone contrast (flat vs falling) – a phonological contrast in which the two most important cues –F0 contour and syllable duration – are supra-segmental features. If feature type is more predictive of imitation, both F0 and duration of Mandarin tones should be well imitated by native speakers. However, if phonological contrast is more predictive than feature type in imitation, we should observe limitations on imitation. We may also observe different patterns of F0 and duration imitation on account of their different importance to the phonological contrast.

[1] In this dissertation, the term 'feature' is used in a generic sense to refer to specific acoustic dimensions such as F0, duration, voice onset time and vowel formants. It is not used in the sense of phonological feature.

2.1.1 Phonetic imitation

Phonetic imitation in general refers to the phenomenon that a speaker adjusts their speech production towards the interlocutor, resulting in an increased similarity between the speakers. The concept was proposed for describing patterns of phonetic convergence in interactive conversations (Coupland et al., 1991). On the other hand, it can also represent noninteractive psycho-linguistic experiments (see Pardo et al. (2017) for a review). There are in general two types of noninteractive tasks. One is spontaneous imitation (also called subconscious, unintentional imitation, or voluntary imitation), where participants are not aware of having modified their habitual production of the content (e.g. Babel et al., 2011). Experiments of this type are also called convergence, accommodation or shadowing experiments. In these experiments participants may be asked to repeat what they hear as quickly as possible (also called close or immediate shadowing) or after a delay (e.g. Goldinger, 1998) or in a production block after an exposure block (e.g. Nielsen, 2011). The other type is explicit imitation (also called forced or intentional imitation), where the participants are explicitly instructed to imitate the voice they heard as faithfully as possible (e.g. Schertz, Adil & Kravchuck, 2023). Studies comparing the two types of imitation found that a greater imitation effect was observed in explicit imitation than in spontaneous imitation (Babel et al., 2011; Dufour & Nguyen, 2013; Garnier et al., 2013), suggesting that awareness of the task increased the degree of imitation. Additionally, we might expect social effects to play less of a role in explicit imitation, hence explicit imitation has been used in studies testing linguistic effects. In this study, we also use the explicit imitation paradigm to directly examine how phonetic imitation is affected by the non-social factors we are interested in. To distinguish these two paradigms, we will follow Pardo et al. 2017 in grouping together close shadowing and exposure tasks and

describe them as 'shadowing' to represent spontaneous imitation experiments. We will use 'imitation' to represent explicit imitation experiments, and 'phonetic imitation' to describe both the explicit and the implicit imitations.

2.1.2 Phonological effects in phonetic imitation

Phonetic imitation is affected by many factors including individual-level effects such as auditory processing manner (Postma-Nilsenová & Postma, 2013), social effects (Coupland et al., 1991; Pardo, 2006) such as the subjective attitude to the talker (Lin et al., 2021; Yu et al., 2013) and the gender of the speaker and talker (Pardo et al., 2017). Furthermore, phonetic imitation is affected by linguistic factors such as word frequency (Goldinger, 1998; Nye & Fowler, 2003; Walker & Campbell-Kibler, 2015) and, of special interest to the current study, phonological effects – wherein imitations are influenced by phonological categories.

In general there are thought to be three phases involved in imitation: perception of acoustic cues, storage in memory, and reproduction (Flege & Eefting, 1988). On the one hand, phonological category effects could come from early phases. Phonological contrasts may warp listeners' perceptual space towards categories – at least for some tasks (Iverson & Kuhl, 1995; Liberman et al., 1957; but see also McMurray et al., 2008). Hence, non-linear acoustic to category mapping could result in non-linear imitation. On the other hand, phonological mediation could also come from a later phase. Category effects are stronger with memory delay, showing that non-phonemic information may be lost in memory (Pisoni, 1973). The production process itself could also produce non-linearities for example because certain sounds have been produced more often. In sum, the source of the phonological effect in imitation is still unclear and all three phrases could be implicated.

The phonological mediation effect has been confirmed in many studies. In an early study, Chistovich et al. (1966) investigated the imitation of an /a-ɛ-i/ vowel continuum by a speaker of Russian. The first three formants were manipulated together in 12 steps. In separate blocks the participant was asked to either repeat the stimulus as quickly as possible or to mimic as closely as possible after hearing it. Results showed that formant frequencies were reproduced in a non-continuous way – that is the distribution of reproduced formant frequencies clustered around several modes rather than the flat distribution of the stimuli. This was true for both tasks. The authors concluded that categorization was automatically involved in imitation. Although Chistovich et al. (1966) analyzed only one participant, the phonological mediation effect was confirmed in subsequent vowel continuum imitation experiments (Kent, 1973; Repp, 1984; Alivuotila et al., 2007).

Several studies of voice onset time (VOT) imitation have found similar results. Flege and Eefting (1988) investigated VOT imitation for three groups of adult participants (although there were also child participants, we do not discuss them here): monolingual Spanish speakers, monolingual English speakers and Spanish-English bilinguals. They used a continuum from /da/ to /ta/ by varying VOT from -60 ms to 90 ms in 10 ms steps. Results showed that all three groups imitated the VOTs in a categorical way – both monolingual groups imitated the VOTs into two modes and the stimuli between the modes were imitated with larger variations. In particular, the two modes of Spanish monolinguals were pre-voiced (-88ms) and short lag (18ms) while those of the English monolinguals were short lag (22ms) and long lag (70ms), reflecting the effect of native phonology. The bilingual group imitated the VOTs according to all three modes. Mitterer and Ernestus (2008) also found that phonetic imitation is phonologically-dominant based on their findings from the shadowing of Dutch

stops. They manipulated the voiced stops /b/ and /d/ into three levels: no pre-voicing, canonical partial pre-voicing and lengthened pre-voicing. Results showed that the difference between no pre-voicing versus pre-voicing was significant in the shadowing, but the difference between the normal amount of pre-voicing and lengthened pre-voicing, which was more phonetic than phonological, was not significant in the shadowing. This result indicates that the phonetic detail is more likely to be shadowed if it's phonologically relevant. Finally, Nielsen (2011) manipulated the VOT of English /p/ into three levels: normal, lengthened and shortened, and found that the lengthened VOT of /p/ was imitated by native English participants but the shortened VOT was not. Schertz, Adil and Kravchuck (2023) made a similar comparison but with explicit imitation. In contrast to the previous study, they found imitation of both lengthened and shortened VOT. However, there seemed to be qualitative differences between the two conditions with a small but consistent shift for imitations of the shortened VOT and larger but inconsistent shifts in response to the lengthened VOTs. The authors propose that a constraint for contrast preservation may be partially responsible for the differences in imitation of the two conditions. Together these studies point to an important role for phonological categories in shaping imitation of stop consonants.

Furthermore, other studies have investigated the phonological effect not only through the primary cue of the phonological contrast, but also by examining non-primary cues. Here we use 'primary' and 'non-primary' cues to refer to the extent to which listeners rely on each cue to distinguish the phonological categories, with the primary cues having the greatest influence (see Schertz & Clare, 2019 for discussion). However, this is not intended to be a binary or qualitative distinction as cue weight is a continuous value that is thought to be closely related to how well the cues distinguish the categories in production (Clayards et al.

2008, Holt & Lotto, 2006, Toscano & McMurray, 2010). Kwon (2019) compared Seoul Korean speakers' shadowing of the primary (F0 of a post stop vowel) and non-primary (VOT) cue to aspirated stops. Results showed that exposure to an enhanced non-primary cue (longer VOT) gave rise to enhanced production of both the non-primary cue (longer VOT) and the primary cue (higher vowel F0) in the post-exposure test. However, the enhanced primary cue condition (higher vowel F0) gave rise to only the enhancement of the primary cue in the post-shadowing test (note however that the influence of the enhanced cue on shadowed productions did not show this asymmetry). This suggests that imitation faithfulness depends on the primacy of the cue to the phonological contrast. In addition, Kim and Clayards (2019) investigated the imitation of an /ɛ/-/æ/ continuum (from *head* to *had*) in English which varied in both formants, the primary cue, and vowel duration, the non-primary cue. They found that for formants, the imitation of the ambiguous step was more variable than the prototypical steps, and tended towards the prototypes. However, the imitation of duration did not show this nonlinear pattern and each duration step was well imitated. The results align with Kwon's (2019) findings in that imitation performance is asymmetrical between primary and non-primary cues. The different imitative patterns between formants and duration could result from their different roles in signalling the contrast – formant was non-linearly imitated because it serves as the primary cue to /ɛ/-/æ/ and is thus most influenced by the phonological mediation effect. Duration was linearly imitated because it is a non-primary cue to the contrast and therefore less susceptible to phonological warping. However, as discussed in Kim and Clayards (2019), other possibilities also exist. Among others, they suggest a reason for the linear imitation of duration and non-linear imitation of formants could be because duration is a 'supra-segmental' feature and formants are 'segmental' features.

2.1.3 Supra-segmental vs Segmental features in imitation

In fact, features such as F0 and duration (or speech rate) have typically been found to be 'easy to imitate' or 'imitated linearly' in previous studies. Earlier studies investigating social factors of spontaneous imitation (i.e., convergence) and studies investigating dialect or speaker characteristic imitation often examined longer stretches of speech such as sentence or conversations. They tended to find that F0, intensity and speech rate converged to the speaker (D'Imperio & Sneed, 2015; Giles et al., 1991; Pardo, 2010; Zetterholm, 2007), whereas the convergence of formants was less consistent (Babel, 2009; Pardo, 2010). Some continuum imitation studies also found that the manipulated F0 range between some intonational pairs (e.g., H* vs L+H*) was imitated linearly (Dilley, 2010; Michalsky, 2015). These findings led other researchers to hypothesize that 'supra-segmental features' were easier to imitate than 'segmental features' (Garnier et al., 2013; Kim & Clayards, 2019). Specifically, Sato et al. (2013) investigated imitation of F0 and F1 in three English vowels /i, e, ɛ/. Results showed that F0 was imitated more than F1 in both spontaneous and explicit imitation tasks. This direct comparison between 'segmental' and 'suprasegmental' cues provides support for the hypothesis.

Other results also support this hypothesis. Wagner et al. (2021) observed native Dutch speakers spontaneously converge to non-native Dutch speech in a disguised shadowing task. In particular, most of the features showing convergence were 'supra-segmental' ones such as f0 and vowel duration. Other studies also found this supra-segmental advantage in second language (L2) imitation. Hao and de Jong (2016) investigated the L2 imitation of two groups of participants, native English speakers' imitation of Mandarin tones (cued by F0) and native Korean speakers' imitation of English stops (cued by VOT). Results showed that native

English speakers performed better in the L2 Mandarin tone imitation task than in the tone production and perception tasks. In contrast, the accuracy of Korean participants' L2 English stop imitation was worse than their production or perception accuracies. This was true even though the native English imitators had fewer years of L2 learning experience than the native Korean imitators. This seems to confirm that imitation of the non-native tones was easier than imitation of non-native stops.

Given these results, it is plausible that the linear imitation of duration observed in Kim and Clayards (2019) was because duration is a 'supra-segmental feature', and similarly, the non-linear imitation of formants was due to their 'segmental feature' status. However, there are limitations to this interpretation. First, the concepts of 'supra-segmental feature' and 'segmental feature' are somewhat vague. From the view of phonology, a linguistic phenomenon in spoken language can be classified as either segmental (e.g., vowels and consonants) or suprasegmental (e.g., presence or absence of a pitch accent) (e.g., Lehiste 1970; Cutler, Dahan & Van Donselaar, 1997), correlating with the widely accepted notion of segmental and suprasegmental tiers (Goldsmith, 1990). However, it is harder to classify features as segmental or suprasegmental. On the one hand, features including f0, duration, intensity and pause are commonly measured for suprasegmental phenomena, and thus are labeled 'suprasegmental features' in many studies (Cutler et al., 1997). In other cases, the same acoustic variables may serve as cues to a segmental phenomenon, such as when differences in F0 following a stop can serve as a cue to stop voicing category (Whalen et al., 1993). With respect to the duration feature investigated in Kim and Clayards (2019), it is difficult to say whether it should be characterized as a 'supra-segmental feature', given that

duration here is a cue to a vowel contrast (/ɛ/-/æ/) rather than a supra-segmental phenomenon.

Secondly, it is not the case that 'supra-segmental features' are always easily imitated. Pierrehumbert and Steele (1989) manipulated the F0 peak position of a rise-fall-rise intonation pattern in English – a suprasegmental phenomenon – in an explicit imitation experiment. Results showed that participants imitated early and late peaks well, but were not able to accurately imitate the peaks which were intermediate. They concluded that there was a boundary within the F0 peak continuum, and that the timing of the F0 peak was interpreted as a binary distinction (possibly corresponding to two categories: L*+H and L+H*) rather than a continuous variable. This study showed that the imitation of F0 peak position, which could be seen as a 'supra-segmental feature' in a broad sense, was mediated by the proposed L*+H vs L+H* distinction, further calling into question the 'supra-segmental feature' superiority hypothesis.

2.1.4 The current study

In brief, so far it's unclear what drives the differing imitation patterns of formant and duration in Kim and Clayards (2019). Although the cue primacy effect is plausible, one cannot exclude the possibility of the feature type effect (i.e., 'segmental vs supra-segmental features'). The evidence from intonation imitation, as reviewed above, is also mixed. One reason may be that for many sentential intonational contrasts it is unclear whether they are two phonological categories or a continuous variation in nature (Ladd & Morton, 1997; Roessig et al., 2019; Dilley, 2010). For example, there are debates whether 'intonational emphasis' should be regarded as categorical or continuous (Ladd & Morton, 1997). As reviewed earlier,

the H* vs L+H* pitch accent pair, which is supposed to be a typical realization of the non-emphasis vs emphasis contrast, was found to be imitated linearly (Dilley, 2010), suggesting that it is not subjected to the phonological effect in imitation thus is not able to be used to examine the two accounts. Lexical stress is another example of a supra-segmental phenomena which can distinguish between words. A recent study investigated imitation of F0 and duration as cues to lexical stress (MacLeod and Di Lonardo Burr, 2022) in Spanish. They found that the model talkers used F0 more than duration, imitators on average adapted their use of F0 and duration towards the model talkers, and that F0 changed more than duration. However because the study did not give imitators a continuum of values including ambiguous ones, we can't tell if their imitations were subject to categorization effects.

The imitation of Mandarin lexical tones allows us to address this question more directly. Lexical tones unarguably play a role in categorizing words (Wang, 1976; Peng et al., 2010) and should therefore be subject to categorical mediation. At the same time, lexical tones are clearly supra-segmental phenomena and are cued by both F0 and duration (see the tone space in Figure 6 (a) in Xu, 2005; Blicher et al., 1990; Yang, 1989; Zhu et al., 2016, but also see Feng & Peng, 2018; Wang & Peng, 2012; Wang et al., 2017 for results that did not observe a duration effect in perception. Possible reasons for the discrepancy in the duration role will be discussed in Section 2.5.2), two cues known to be well imitated in many studies.

Thus, in Mandarin lexical tones we have a case where both the primary (F0) and the secondary (duration) cues are supra-segmental features allowing us to test the two potential explanations. The feature type account predicts good imitation performance, i.e., linear imitation, for both cues. The cue primacy account predicts that the more a cue contributes to a categorical percept (i.e. the more important it is to a contrast), the more categorically it is

imitated, hence F0 should be more categorically-imitated and duration should be more linearly imitated. Two hypotheses can be made:

• Hypothesis A: If feature type dominates phonetic imitation, both the F0 pattern and the duration of the tone should be linearly-imitated.

• Hypothesis B: If cue primacy to the phonological contrast dominates phonetic imitation, duration will be linearly imitated while F0 will be categorically-imitated.

To the best of our knowledge, only one other study has investigated the imitation mechanism of lexical tones. Lin et al. (2021) investigated how the two Cantonese level tones that are undergoing a merger were imitated by native speakers. The current study will be the first to investigate the imitation mechanism of contour tones.

In our study we chose the (high) flat and falling tone pair for our tonal continua. The falling tone is the shortest tone intrinsically making it a good candidate for studying duration as a secondary cue (Xu, 2005). Secondly, although the low tone is the longest, we did not choose it since the cues of the low tone include other features such as potential creaky voice, the F0 valley timing, and the F0 valley duration (Jongman et al., 2006). Furthermore, the tonal target for the low tone varies in isolation and speech in context.

In summary, the primary research question addressed by this study is the extent to which the F0 contour and duration of tonal stimuli are imitated in a linear or categorical way, to better understand imitation asymmetries. Two experiments were conducted on manipulated continua varying between the high flat and the falling contours. Continua were created by manipulating F0 falling range and syllable duration. The first experiment was a perception task to confirm the role of duration on the flat-falling tonal contrast and to identify the

perceptual categorization of each stimulus. If imitation is strongly driven by categories, the category may depend on the combination of duration and F0 range. The second experiment was an imitation task in which the main research questions will be investigated.

2.2 Material

All data, analysis scripts, experimental materials and preregistration are available at (https://osf.io/7c2tf/).

2.2.1 Stimuli

2.2.1.1 The target word

The syllable 'ba'[pa] was used to create the tonal continua. After considering different consonants (and a non-consonant choice), we chose 'b' for two reasons: (a) It typically does not exhibit syllable-initial glottalization (as can happen with vowel initial syllables), and gives rise to weaker microprosodic F0 perturbations in the following vowel than the other consonants we tested in pilot work. Thus using 'b' can allow us to estimate more precisely the onset F0 values in the tonal trajectories than other consonants. (b) The duration of 'b' is short and relatively constant (compared to e.g. fricatives or nasals), such that it would not lead to additional effects on the production of duration.

Pilot work showed that the target syllable 'ba' is often produced with severe creaky voice at the vowel offset when it is read in isolation. Since creaky voice complicates the interpretation of extracted F0, we placed the target syllable in the initial position of the tri-syllabic word '"ba"的声 [pa.de.ʂɤŋ], PINYIN: ba de shēng, 'the sound of 'ba''. In order to

minimize the amount of anticipatory coarticulation between the tonal targets of 'ba' and those of succeeding syllables, we used 'de', a function word with neutral tone, immediately after 'ba'. This tri-syllabic word was used in the perception and imitation experiments.

2.2.1.2 The continua

We inspected the spectrograms of a list of recorded words and phrases (30 altogether) produced by five female native speakers of Mandarin. We chose one of them as the speaker for generating our stimuli based on the overall F0 height, the absence of creaky voice, and a smooth amplitude envelope. This speaker recorded five repetitions of the words 'bā', 'bà', 'bā de shēng', and 'bà de shēng'. The average duration and F0 values of each syllable are listed in Table 2.1. The F0 values for non-falling tones (i.e., other than bà) are the average F0 of the vowel. For falling tones (i.e., bà), the F0 is represented by onset and offset values.

Table 2.1:Average duration and F0 features from five repetitions by the speaker used to set stimulus values. For bā, de and shēng, the F0 values are mean F0

	'bā'	'bà'		'bā de shēng'			'bà de shēng'			
				bā	de	shēng	bà		de	shēng
Duration (ms)	368	275		233	189	493	227		185	443
F0 (Hz)	244	Onset 273	Offset 162	270	236	217	Onset 299	Offset 223	204	228

We chose one 'bā de shēng' repetition with the least perturbation and smoothest amplitude envelope from the five productions as the source sound. When generating the continua, we manipulated the target syllable 'bā' and the carrier 'de shēng' in the tri-syllabic word separately, and then spliced them together in order to create the different stimuli.

The 'ba' continua varied along two dimensions, F0 range and duration. The F0 patterns of the continuum endpoints 'bā' and 'bà' were set as the average F0 values in Table 2.1. (270 Hertz (Hz) for the flat tone, and 299-223 Hz for the falling tone, range 76 Hz). We manipulated the F0 dimension by evenly interpolating 7 intermediate onset and offset steps between these two endpoints to create 7 ranges, as shown in Table 2.1. Values for ranges (in both Hz and semitone (st)), onsets and offsets (in Hz) are listed in Table 2.2.

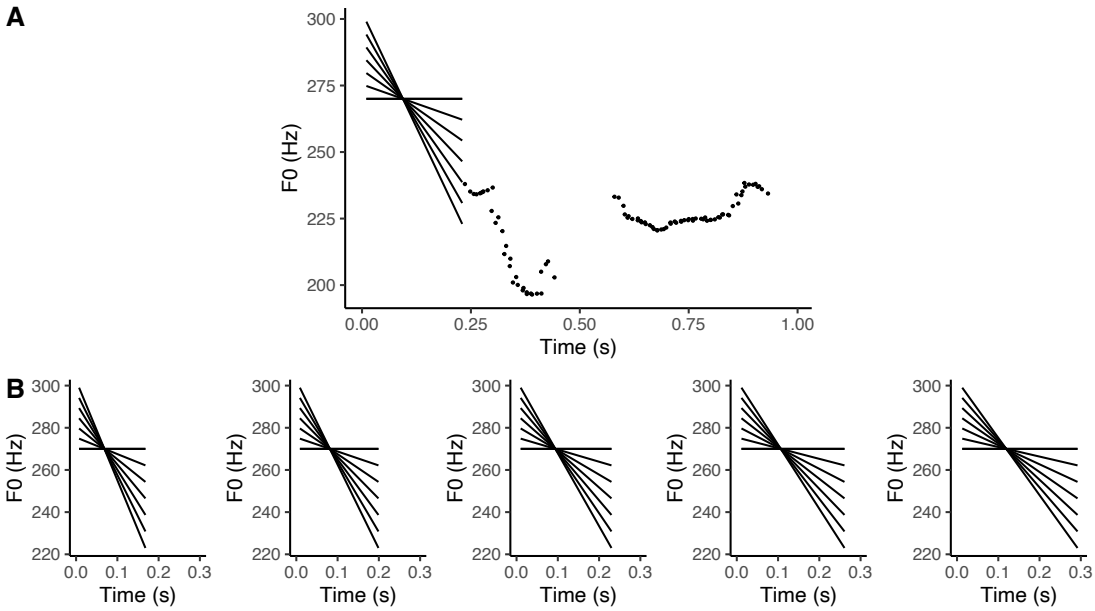


Figure 2.1: Schematic of the experimental stimuli. Panel A: One set of F0 range steps for the middle (230 ms) duration step along with the average pitch track of the last two syllables from the carrier 'de shēng'. Panel B: all duration by F0 steps of the target syllable 'ba'.

Table 2.2: Onset, Offset and F0 falling range values for each F0 range step in the stimuli

Step	1	2	3	4	5	6	7
Onset (Hz)	270	262	254	247	239	231	223
Offset (Hz)	270	275	270	285	289	294	299
F0 Range (Hz)	0	13	25	38	51	63	76
F0 Range (st)	0	0.8	1.6	2.5	3.3	4.2	5.1

The middle step of the duration continuum was calculated by averaging the durations of 'bā' and 'bà' in the ten tri-syllabic words (230 ms). Note that 230 ms included the 10ms of the stop consonant release. In all the stimuli we manipulated the duration of the vowel only, and the consonant remained 10 ms long in all stimuli. The duration value of the shortest step was set as the shortest duration needed to realize the falling F0 range of 'bà' (76 Hz or 5.1st), which was to make sure that participants were able to realize the F0 falling range in terms of the physiological limitation. The shortest duration was calculated according to the maximum speed of pitch change reported in Xu and Sun (2002). We used the fall velocities for a 4 st falling excursion (-27 st/s) and a 7 st falling excursion (-37 st/s), to interpolate roughly -30 st/s for a 5.1 st fall. The shortest duration step was thus 168 ms. We created a maximum duration that was equally far from the mean (292 ms) and two intermediate steps for a total of 5 steps. The duration continuum thus varied in 5, 31 ms steps (168, 199, 230, 261, 292 ms).

The 7 F0 range steps and the 5 duration steps were combined to create 35 stimuli. All manipulations of the original recording were done in Praat (Boersma, 2006) by replacing the duration and pitch tiers and resynthesizing the new stimuli using the PSOLA algorithm.

The tonal continua were then spliced onto the manipulated carrier from the original recording. To avoid any bias to the perception of the tonal continua, we neutralized the duration and F0 contour of the carrier 'de shēng'. We set the duration of 'de' and 'shēng' as the average across the 10 productions (5 from 'bā de shēng' and 5 from 'bà de shēng') respectively. The F0 contour of 'de shēng' was manipulated as the average F0 shape of the 10 'de shēng' productions. The splicing was done manually by the first author, at zero crossings to avoid popping sounds produced by discontinuities in the waveform. Figure 2.1A shows

one full set of 7 F0 range stimuli for the middle duration step (230 ms) along with the pitch track of the carrier sentence used for all stimuli. Figure 2.1B shows all duration by F0 steps of the target syllable 'ba'.

We note that previous studies have manipulated F0 contour in contour tones (i.e., the rising or falling tones) either in terms of F0 range, which varies how much the f0 falls/rises, or F0 slope, which varies how fast the f0 falls/rises. There is some debate about which of these is the best feature for the contour tones. Because slope and range are related to each other through duration, it is impossible to manipulate all three independently. We chose F0 range but also performed planned additional analyses on our data by re-coding trials according to F0 slope. We then compared the imitation distributions for F0 slope and F0 range. We had hoped this might shed light on which was a better cue to the contrasts, however we did not find any conclusive differences between slope and range. The additional analyses can be found in the OSF repository (https://osf.io/7c2tf/).

2.3 Experiment 1: perception

Our first goal was to determine how the 35 stimuli were categorized by native listeners. As reviewed in Section 2.1, although some studies found that duration was a secondary cue to the Mandarin flat-falling tonal contrast, others did not. The discrepancy could result from different stimuli setups across studies, thus it is important to establish the role of duration with our stimuli. It will also allow us to see how each stimulus is categorized on average as either the flat or falling contour. Specifically, for each duration step, we will estimate the F0 range at which perception crosses over 50% falling contour responses.

2.3.1 Participants

We recruited 15 native Mandarin participants (mean age: 27, standard deviation: 3.2, M: 5 F: 10) through Prolific, an online recruitment platform. All participants were required to be native speakers of 'Chinese' as reported to Prolific. All participants also reported that they grew up in China. Seven of them grew up in the northern part of China, and the rest grew up in the southern part of China. The online experiment was set up using the prosodylab experimenter (Wagner, 2021), a set of javascripts building on jsPsych (De Leeuw, 2015).

2.3.2 Procedure

This experiment used a two-alternative forced choice method. In each trial, one of the 35 stimuli was played and two choices "ba的声" (the sound of bā) and "bà 的声" (the sound of bà) were given on the screen. Participants were instructed to choose one of the options by pressing one of two number keys. There were 5 repetitions for each stimulus and the stimuli were presented in random order. The experiment took approximately 15 minutes.

2.3.3 Results

In this experiment we were interested in whether duration affected the perception of the flat-falling contrast. We expect 'falling' responses to increase as the F0 range increases. If duration plays a role in perceiving the flat-falling contrast, we expect duration step to also influence the likelihood of a 'falling' response. In particular, since the falling tone is intrinsically shorter than the flat tone, shorter duration should lead to more falling tone responses.

Figure 2.2 shows that 'falling' responses increases with larger F0 ranges (higher step number) as expected. We also see that shorter durations lead to more falling tone responses.

This indicates that duration influences the perception of the flat-falling contrast, as found in previous studies. Longer durations are heard more often as the intrinsically longer tone, i.e. the falling tone. It's worth noting that this is not likely due to contour tones being processed in a special way. Previous work on the flat-rising contrast found that when the intrinsically longer tone is the flat tone, longer durations were heard more often as the flat tone (Yang, 1989). This is in line with similar studies on e.g. intrinsic vowel length in English (Hillenbrand et al., 2000).

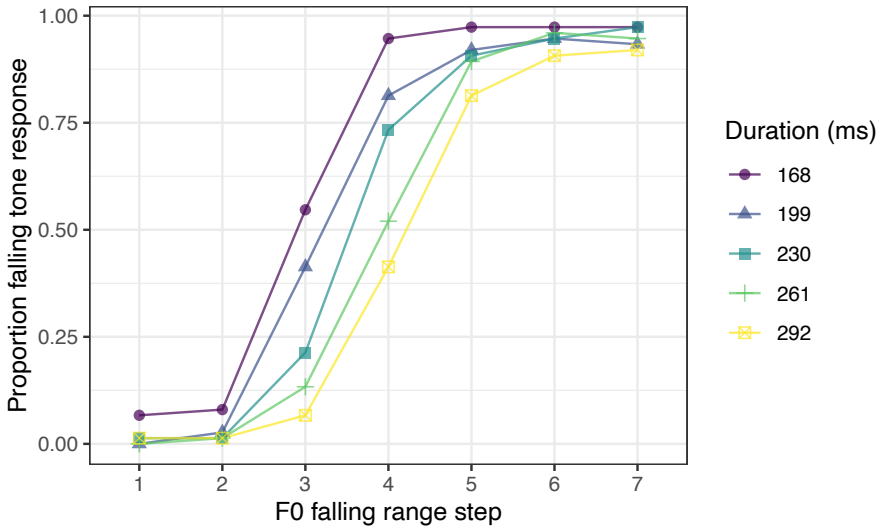


Figure 2.2: Rate of response of the falling tone as a function of the F0 falling range step, grouped by duration step.

These observations were confirmed by fitting a mixed-effect logistic regression model for the response rate of the falling tone. The model was implemented on the R platform (R Core Team, 2013). Duration and F0 range (the step indices rather than measurements) were rescaled and added as fixed effects. By-participant random intercepts and random slopes of both duration and F0 range were used. The model results are shown in Table 2.3. Larger F0 falling range increased the falling tone response ($\beta = 12.22$, se = 0.39, $z = 6.42$, $p < 0.01$) as

expected. Importantly, longer durations decreased the response rate of the falling tone (β = -2.59, se = 0.34, z = -7.64; $p < 0.01$). This result confirms the role of duration in cuing the flat-falling tonal contrast. We can also see that the coefficient for F0 range is much bigger than the coefficient for duration. Since both variables were scaled, the magnitude of these coefficients is an indicator of their role in signaling the contrast, confirming that duration plays a smaller role. We used the fitted regression equation to estimate the point along the 7 step F0 range continuum where responses were 50% as a measure of the category boundary. For example, the estimated boundary for the shortest duration step was F0 range step 3 (25 Hz) while the estimated boundary for the longest duration was just past step 4 (38 Hz) at 4.2 steps. The estimated boundaries (in steps) for each duration were: 168 ms, 3.0; 199 ms, 3.3; 230 ms, 3.6; 261 ms, 3.9; 292 ms, 4.2.

Table 2.3: Summary of the mixed-effect logistic regression model for the response rate of the falling tone

	Estimate	Std.Error	Z value	p
(Intercept)	1.25	0.39	3.20	< 0.01
F0 range	12.22	1.90	6.42	< 0.01
Duration	-2.59	0.34	-7.63	< 0.01

2.4 Experiment 2: imitation

2.4.1 Participants

We recruited 18 new native Mandarin participants (mean age: 25, standard deviation: 3, gender-balanced) through Prolific. They all reported that Mandarin Chinese is their first language or dominant language. In addition, the first author examined their nativeness by listening to their production of the short passage. One participant was judged to be not native-like hence was excluded and replaced. Fifteen of the participants grew up in China, with three of them being from Beijing and the rest from other southern provinces. Three participants reported growing up in Canada or Malaysia. The online experiment was also set up using the prosodylab experimenter (Wagner, 2021).

2.4.2 Procedure

This study used an explicit imitation paradigm. There were three blocks in the experiment. The first was the baseline block, where the participants naturally produced 'ba' with the flat and the falling tones (designated by the PINYIN form 'bā' and 'bà' on the screen) in isolation, 5 times each. This block was to show the participants the 'bā' and 'bà' categories ahead of the imitation task in which no orthography would be shown on the screen, and to collect participants' baseline production of the two tones. The second block was a reading block. Participants were instructed to read a short 160-syllable passage 'The North Wind and The Sun' (in Mandarin) after a silent reading warm up. We used this block to assess the Mandarin proficiency of the participants, and to estimate their speaking F0 range. The speaking F0 range was used to improve pitch tracking and to normalize speakers F0. Since tonal targets are not absolute F0 values but relative pitch heights, we used normalized F0 within the speaker's F0

range when comparing the pitch trajectories. According to Baken and Orlikoff (2000), in English, a speaker's speaking F0 range can be estimated from 3-4 sentences. To be more precise, in this experiment we used a short passage. The third block was the imitation task. The participants were instructed to repeat what they heard "越接近话者的发音方式越好" ("the closer to the way the speaker produced it the better"). There were 35 stimuli and 5 repetitions for each.

2.4.3 Results

Target syllable boundaries were first automatically aligned using the Montreal Forced Aligner (McAuliffe et al., 2017), and then manually checked and corrected by the first author. Syllable onset was marked as the onset of vocal fold vibration and the offset was marked at the time that energy started to drop sharply. Syllable duration was measured between these two points.

Praat (Boersma, 2006) was used for the F0 extraction. When extracting F0 values, we used the following procedure to minimize F0 tracking errors. We first analyzed the distribution of F0 values in each participant's reading passage. We computed the mean F0 in Hz (50% quantile) as well as the 25% and 75% quantile for each participant. Second, we calculated each participant's speaking F0 range by the two formulas proposed by De Looze and Hirst (2008): ceiling = $F0_{(75\%\ quantile)}$ * 1.5; floor = $F0_{(25\%\ quantile)}$ * 0.75. Third, for each speaker, their individual ceiling and floor values were used as parameters in Praat to extract the F0s for each of their imitations. In each syllable we extracted 10 F0 values, each of which was the mean F0 within every 10% of the syllable. Roughly compared to the default settings, we found that this method worked well in reducing F0 tracking errors.

To normalize the F0 ranges across speakers, we then transformed Hz to st (semitones) using the formula in (1), in which F0ref is the participant's mean F0 in the reading passage. F0 range was calculated in the following steps: First, the maximum F0 in the first 50% of the target syllable was obtained, denoted as $F0_{Max\_1stHalf}$. According to visualization of the pitch tracks and experience from previous studies (Wang et al., 2020; Xu, 2005), the Max F0 of a falling tone in general happens in the first half of the syllable. Second, the minimum F0 was calculated as the last F0 extracted, i.e., the mean F0 in the last 10% of the syllable, denoted as $F0_{Min\_final}$. In cases where the F0 of last 10% was not extracted successfully, we used the preceding F0 measurement, as the $F0_{Min\_ final}$. Finally, the F0 range was calculated as $F0_{Max\_1stHalf} - F0_{Min\_final}$.

$$\boldsymbol{st = 12 * log_2(\frac{F0}{F0_{ref}})} \quad (1)$$

2.4.3.1 Baseline Tone Productions

The full data set included 180 productions (90 for each tone) from the 18 speakers which were produced in the baseline task before the imitation task. Twelve productions were excluded due to inability to extract F0 in the last 40% of the syllable for reasons such as creaky voice. Two productions were further excluded for either being shorter than 100 ms or showing an F0 increase of larger than 50 Hz (when a decrease was expected). The remaining 166 productions (92%) were used for analysis. Table 2.4 summarizes the F0 range and durations of the two tones. These are in line with the values observed with our imitation model talker as summarized in Table 2.1.

Table 2.4: Mean and standard deviation (SD) of F0 range and duration in the baseline tone productions

Tone type	Duration (ms)		F0 range (st)	
	Mean	SD	Mean	SD
Flat	364	91	0.36	1.12
Falling	241	76	4.88	2.76

We also wanted to quantify the role of each of the two cues in signaling the tone category by examining the overlap of distributions. Figure 2.3 illustrates the density distribution of F0 range (after normalization and conversion to st) and duration for the two tones produced in isolation. The distribution of duration for the two tones shows a substantial overlap, while the F0 range distributions are relatively more distinct. We quantified the separation of the two categories using Cohen's d (Cohen, 1988) which is calculated using the means and variations of the two distributions. A larger Cohen's d indicates a larger separation. The Cohen's d values are 2.1 for F0 range and 1.5 for duration. This is in line with our perception results, supporting the claim that F0 range is a more important cue than duration in signaling the flat-falling tonal contrast.

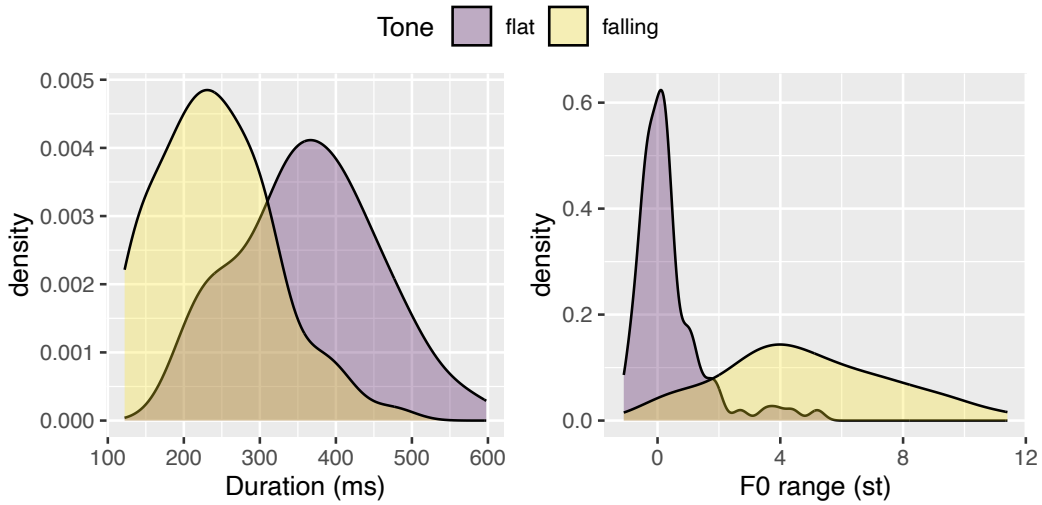


Figure 2.3: Density distributions of duration (left panel) and F0 range (right panel) for the flat tone and the falling tone in isolated productions, pooled from all speakers.

2.4.3.2 Imitation of duration and F0 range

The full data set included 3150 (18*35*5) imitations of 'ba' from the 18 speakers. 27 recordings were removed because of recording errors or missing data, leaving 3123 valid imitations. 33 were further excluded as neither the 10th or the 9th F0 was extracted successfully. Four were excluded since none of the first 5 F0s was extracted successfully. Another sample was excluded because the F0 range was -87.3 Hz which was an unexpected extreme rising contour. The final data set included the remaining 3085 imitations (98%) used for data analysis.

Figure 2.4 shows F0 trajectories of the imitations. In the figure, the pitch trajectories have been time-aligned to the end of the target syllable 'ba', as indicated by the dashed line. It can be observed in Figure 2.4 that the range of F0 fall in the imitation increases as the F0 range in the stimulus increases. The following analyses focus on the target syllables only.

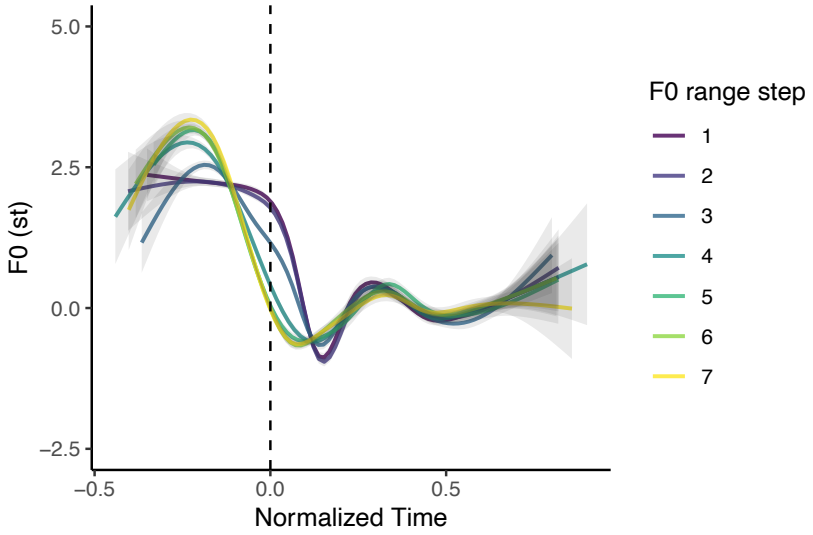


Figure 2.4: F0 tracking of imitation, grouped by the F0 range step of the stimuli. Each line is smoothed using a generalized additive model fit to the empirical data for visualization purposes. Shaded areas are the 95% confidence intervals of the smoothed lines. Step 1 is the flat tone and step 7 is the falling tone. The dashed line indicates the end of the target syllable 'ba'.

Figure 2.5 shows density plots of the imitations (Left: duration, Right: F0 range) as a function of the target values in the stimuli (the target F0 ranges were transformed to st). We observed a clear difference in the imitation of the two features: the imitation of duration was accurate, and the relationship between the duration of the stimulus and the duration of the imitation was roughly linear, consistent with what was observed in Kim and Clayards (2019) for vowel duration. In contrast, the relationship between the F0 range of the stimulus and the F0 range of the imitations is non-linear. If we pool all F0 range values together, there seems to be a bi-modal distribution, with the first mode being close to zero and the second being close to 4 st. These two modes seem to correspond to prototypical F0 falling ranges for the flat and falling categories.

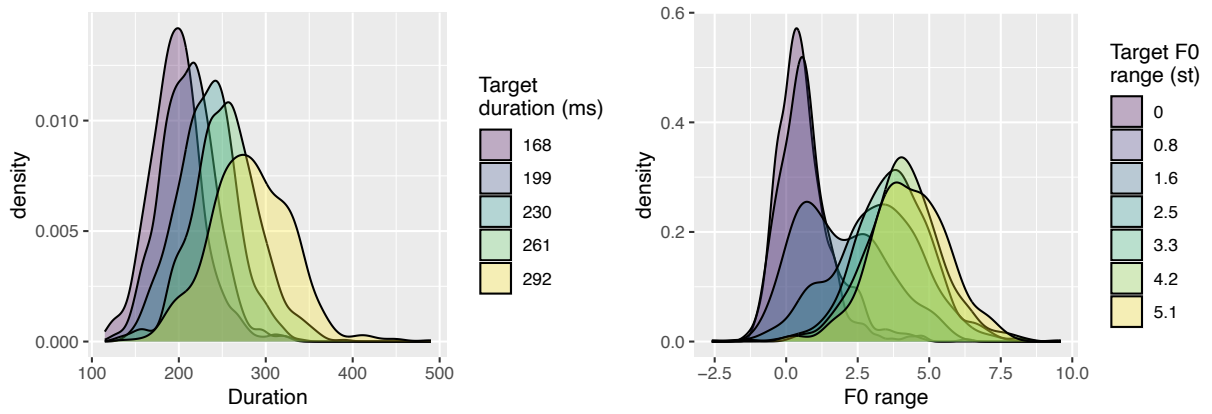


Figure 2.5: Density distribution of the imitated duration (on the left) and F0 range (on the right) for each step of the continua. Colors indicate the levels of the target stimuli.

In Experiment 1 we investigated the effect of duration on the perception of the flat-falling contrast and found that the category boundary along the F0 range dimension shifted as duration increased. In Experiment 2, we are also interested in how duration affects the imitation of F0 range, and whether there is any correspondence to the perceptual result.

In Figure 2.6, each of the 5 panels shows the F0 imitation of a duration step. The effect of duration can be observed on the third step (1.6 st) and the fourth step (2.5 st) of the F0 range continuum across the 5 durations. As the duration of the stimulus increases, the variation in the imitation of the 1.6 st stimuli decreases, and the center of the distribution moves towards the flat tone category. This seems to indicate that the 1.6 st stimuli are shifting from being ambiguous to being imitated as the flat tone, i.e. the tonal boundary between the tones is shifting from around a 1.6 st drop to a larger drop, across the 5 durations. The imitation of the 2.5 st stimuli also varies with duration. When the stimuli are short (168 ms), the center of the imitations is close to the falling tone category. As the duration increases, the distribution of the imitations gradually moves towards the flat tone category. These results are consistent with what was found in the perceptual experiment: longer stimuli are more

likely to be categorized as the flat tone, and the perceptual boundary between flat and falling stimuli moves from step 3 (1.6 st) to step 4.2 (> 2.5 st) as duration increase from 168 to 292 ms.

Figure 2.6 also lets us make an additional observation that was hinted at in Figure 2.5. Within the perceptual category of the falling tone (168 ms, F0 range steps > 1.6 st; 199-261 ms, F0 range steps > 2.5 st; 292 ms: F0 range steps > 3.3 st), the imitated F0 falling range increases as the F0 target range increases. In other words, falls with larger range are imitated with larger range. This suggests that even though the imitation of F0 range is mediated by phonological contrast, phonetic details were maintained to some extent.

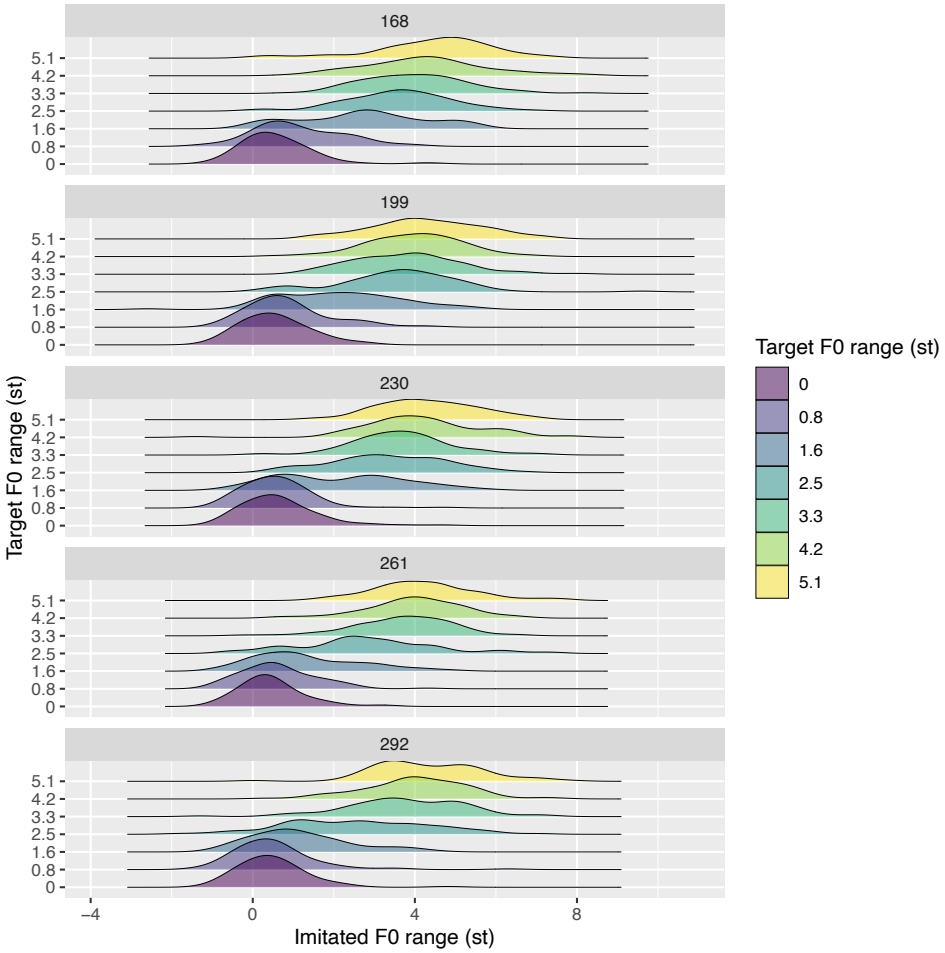


Figure 2.6: The duration effect on F0 range imitation. Each panel shows one level of duration of the stimuli.

To further quantify these observations, we fit a unimodal and a bi-modal model to both distributions. Then we compared which of the two models (bi- or unimodal) was a better fit for the imitation data. If the bi-modal distribution is a better fit, it would indicate that the imitation of the feature is mediated by the tonal contrast whereas if the unimodal distribution is a better fit, this would indicate that the imitation of the feature is less affected by the tonal contrast.

The models were implemented as Bayesian mixed effects models using the *brms* package (Bürkner, 2017). We used a Gaussian-mixture model to build up a bi-modal regression model (with two gaussian mixture components), and a simple Gaussian model for the unimodal regression model using the family parameter of the *brms* package. For a bi-modal model with two mixture components, a mean and sigma (standard deviation) for each component are fit to the data. By using a mixed effects model we can specify a more complex model, e.g. where participant specific means and sigmas are fit using the random effects structure. For the bi-modal model, taking F0 range as the example, the following formulas were used,

$$f0\_range \sim 1 + (1 \mid participant) \quad (2)$$

$$theta1 \sim 1 + target\_f0\_range * target\_duration + (1 + target\_f0\_range \mid participant) \quad (3)$$

$$sigma1 \sim 1 + (1 \mid participant) \quad (4)$$

$$sigma2 \sim 1 + (1 \mid participant) \quad (5)$$

Using the formula in (2), the means of each Gaussian mixture were estimated from *f0_range*. The by-participant random effect allows each imitator to have a different mean for each mixture. Formula (3) fits the mixture probability theta. This allows the model to use each imitation's target F0 range and target duration to predict the probability that the imitation belongs to the first (rather than the second) mixture. We expected the target F0 range to influence the F0 range of the imitation and we also hypothesized, based on the perception results and the observations in Figure 2.6, that target duration could affect F0 range imitation. For completeness we included the interaction between these two factors. To ensure model convergence, only the target F0 range was added as random slope. This allows the assignment of imitations to each mixture component – based on the target F0 range of the stimulus heard – to differ for each participant. Formula (4-5) fit the sigmas of the two mixtures and allow each imitator to have different sigmas. The priors of the Bayesian models were set according to Figure 2.5. For example, the means of the two components in the mixture model of F0 range were taken from normal distributions with means of 0 and 4 respectively and standard deviations of 1 (i.e. *normal* (0, 1) and *normal* (4, 1)). These were chosen to match the modes of the two bumps observed in the empirical data and not overlap. All the priors used in this study are listed in Appendix 2.A.

In terms of the unimodal model, the following formulas were used,

$$\text{f0\_range} \sim 1 + (1 \mid \text{participant}) \tag{6}$$

$$\text{sigma} \sim \text{target\_f0\_range} + (1 + \text{target\_f0\_range} \mid \text{participant}) \tag{7}$$

The formula in (6) is identical to (2), and fits a single mean for the Gaussian with a by-participant random intercept. Formula (7) allows the sigma (i.e., the standard deviation) of

the Gaussian to vary according to the target F0 range. From Figure 2.5, the distribution seems to be skewed if it is considered as a simple Gaussian, and the variation is larger as the target F0 range increases. Therefore we used formula (6) to improve the fit of the model. As in the bimodal model we allowed target F0 range to vary by participant using a random slope. We didn't include duration in this model as we didn't observe a main effect of duration on F0 variance.

The Bayesian models of duration imitation were implemented the same way as F0 range imitation. We should note however, that the mixture model of duration imitation suffered from convergence issues when we used the formulas in (2-5). In order to achieve convergence, we then tested several times by increasing the iteration value and reducing model parameters. The final formulas for the mixture model were,

duration ~ 1 (8)

theta1 ~ target_f0_range* target_duration + (1 | | participant) (9)

sigma1 ~ 1 + (1 | participant) (10)

sigma2 ~ 1 + (1 | participant) (11)

and the formulas for the unimodal model were,

f0_ duration ~ 1 + (1 | participant) (12)

sigma ~ duration_target + (1 + duration_target | participant) (13)

Performance of the unimodal and bi-modal models were compared by means of a leave-one-out (LOO) cross-validation (using the *loo()* function in R) for both F0 range and duration. Results are summarized in Table 2.5. The difference of the expected log pointwise predictive

density (ELPD) between the two models showed statistically strong evidence that the mixture model was a better fit for the F0 range imitation data, and that the unimodal model fit the duration imitation better. The statistical analyses confirm that the imitation of duration is better described as a single distribution of values along the continuum, whereas the imitation of F0 range is better described as a bi-modal distribution.

Table 2.5: The model comparison results (ELPD: log pointwise predictive density. SE: standard deviation)

Fits for F0 range imitation		Fits for duration imitation	
Model	ELPD (SE)	Model	ELPD (SE)
Mixture	**0 (0)**	**Unimodal**	**0 (0)**
Unimodal	-1895.6 (55.7)	Mixture	-501.5 (31.4)

The summary of the mixture model for F0 range is shown in Table 2.6 (note that the sigmas are log transformed in all models). The estimates of the mixture centers are 0.56 st and 3.88 st. From the perceptual results using these stimuli (Experiment 1), the categorical boundary between flat and falling tone varied between the third and fourth step, namely, a fall of between 1.6 and 2.5 st. Hence the mixture with a mean of 3.88 st should indicate a clear falling tone whereas the mixture with a mean of 0.56 st should be within the flat tone category. This confirms our previous observation that the two modes are closely related to the falling tone and the flat tone, respectively. In Table 2.6, target F0 range has a credibly negative effect[2] (β = -13.64, 95% CI: -16.46 – -11.27) on the probability of belonging to the first mixture (theta1). This shows that the larger the target F0 range, the less likely the imitation is to belong

[2] A 95% CI that does not include the value of 0 can be taken to provide compelling evidence for an effect.

to the first mixture, and the more likely to belong to the second mixture, i.e. the falling tone. Target duration has a credibly positive effect on theta1 (β = 2.86, 95% CI: 2.2 – 3.67), indicating that all else being equal, stimuli with longer duration are more likely to be imitated as the flat tone, consistent with the patterns in Figure 2.6. In addition, the intercept of theta1 is credibly negative (β = -2.92, 95% CI: -3.96 – -1.91), showing that on average imitations are less likely to belong to the flat mixture than the falling tone mixture. This suggests that overall more stimuli were imitated as the falling tone than the flat tone.

Table 2.7 shows the summary of the unimodal model of duration imitation. The estimate of the mean of the Gaussian is 0.24 s with a sigma of -3.55. The duration step index has a credibly positive effect (β = 0.42, 95% CI = 0.28 – 0.55) on sigma, indicating that imitations varied more as the stimulus duration increased.

Table 2.6: Summary of the Gaussian mixture model for F0 range imitation (CI: credible Interval)

	Estimate	Est.Error	l-95% CI	u-95% CI
Mean of mixture 1 (Intercept)	0.56	0.15	0.26	0.86
Mean of mixture 2 (Intercept)	3.88	0.19	3.50	4.26
theta1 (Intercept)	-2.92	0.52	-3.96	-1.91
theta1: target F0 range	-13.64	1.32	-16.46	-11.27
theta1: target duration	2.86	0.37	2.20	3.67
theta1: target F0 range*target duration	0.89	0.92	-0.90	2.73
sigma1 (Intercept)	-0.63	0.09	-0.79	-0.45
sigma2 (Intercept)	0.08	0.06	-0.03	0.19

Table 2.7: Summary of the unimodal Gaussian model for duration imitation (CI: confidence Interval)

	Estimate	Est.Error	l-95% CI	u-95% CI
Mean (Intercept)	0.23	0.01	0.21	0.26
sigma (Intercept)	-3.55	0.05	-3.66	-3.44
sigma: target duration	0.42	0.07	0.28	0.55

2.4.3.3 Individual differences

Previous sections revealed that F0 range imitation is mediated by the tonal contrast, and syllable duration affects which tone is imitated. As a result, for the intermediate steps of F0 range, the imitation distribution was wider than the extreme steps. For example, imitation of the third step (1.6 st) in the top three panels of Figure 2.6 has a broad distribution straddling both modes of the overall distribution. This may indicate that some of the imitations were produced as the flat tone while others the falling tone. Could these variations be due to different participants having different boundaries of the tonal contrast? In addition, we observed that within the falling tone category, the different F0 ranges in the stimuli were distinguished in the reproductions to some extent. We wondered if these within-category phonetic details are maintained consistently for all imitators, or if this group data was a mix of more categorical behavior and more linear imitations. Thus, after examining the data, we decided to explore these questions by looking at the F0 range imitation at an individual level. We plotted each individual's data in Figure 2.7 and we also used the individual mu and sigma parameters of the bimodal model fitted above to calculate Cohen's d for each individual where larger Cohen's ds are expected for more categorical imitations. Note that since duration imitation was less relevant to the observed non-linearity on the intermediate F0 steps here,

and due to space limitation, we did not analysis the individual-level differences for duration imitation, but we visualized it in our online materials at https://osf.io/7c2tf/ for readers who are interested in it.

Figure 2.7 shows the distributions of F0 range imitations for 3 representative participants who imitated the F0 ranges categorially (participant 1, Cohen's d = 6.1), less categorically (participant 9, Cohen's d = 3.8), and linearly (participant 17, Cohen's d = 1.9). For participant 1 (left panel), the distribution of imitated F0 range shows two clear modes, indicating that the participant always imitated the F0 ranges as one of the two tones and, within each tone, the F0 fall was consistent. For participant 17 (right panel), on the other hand, it seems that each of the F0 falling steps is reproduced in a more linear way. The participant shows the ability to perceive and reproduce the F0 falling range in the stimuli, with minimal influence of the tonal categories. Finally, in the middle panel, participant 9 imitated the F0 range in a hybrid way: there are two modes overall which aligned to the flat and falling tones (categoricalness); but within the second mode, the F0 falling range can be distinguished as well (linearity). The imitations of all speakers and their Cohen's d values are in shown in Appendix 2.B. Across all data, the pattern of participant 17 turned out to be very rare, and more participants showed the pattern of participant 9, i.e., the hybrid pattern. The results overall seem to show that, for most speakers, reproduction of F0 range is affected by the tonal contrast, but some phonetic detail of the falling tone is also maintained.

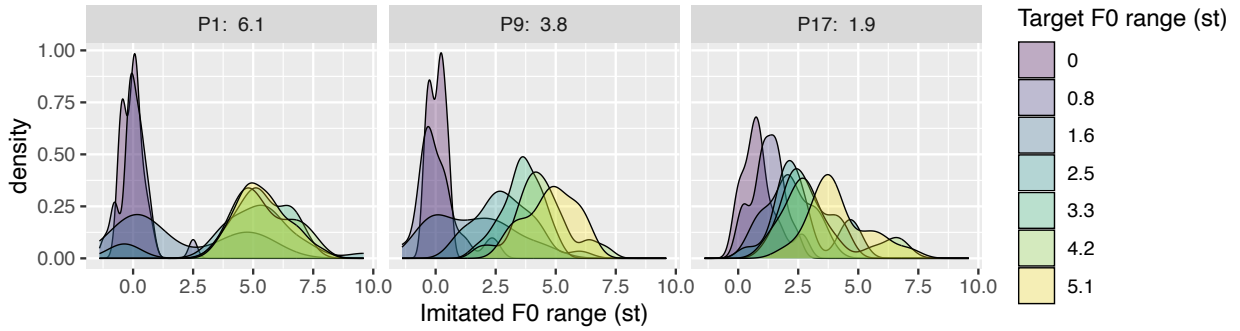


Figure 2.7: Distributions of F0 range imitation and the Cohen's d values of three representative speakers.

2.5 Discussion

This study investigated to what extent the imitation of suprasegmental features is mediated by phonological categories, through perception and imitation experiments of Mandarin flat-falling tonal continua. The continua varied in two suprasegmental dimensions, F0 falling range, which is the primary perceptual cue of the tonal contrast, and duration, which is the less important cue of the tonal contrast. In the perception experiment, the category boundary between the flat and falling tones along the F0 range dimension varies as duration varies, confirming the role of duration in categorization of this contrast. In the imitation experiment, the primary F0 range cue was more categorically imitated, whereas the less important duration cue was linearly imitated. The results revealed that the cue primacy account can predict the phonetic imitation performance better than the feature type account.

2.5.1 Role of phonological mediation in imitation

Results of the current study are in line with the finding that phonetic imitation is mediated by the phonological contrast (Chistovich et al., 1966; Flege & Eefting, 1988; Kim and Clayards,

2019; Nielsen, 2011; Mitterer & Ernestus, 2008). Specifically, we confirmed that the imitation of individual cues depends on their role in the contrast. The primary cue of the flat-falling tonal contrast, F0 range, was mostly imitated categorically, whereas the non-primary cue duration was largely imitated linearly by native speakers, consistent with previous observations that cue-primacy influenced the imitation performance (Kwon, 2019; Kim and Clayards, 2019). Given that F0 is a suprasegmental feature, it suggests that the claim that suprasegmental features are not always imitated linearly.

Furthermore, our results showed that both a primary and non-primary cue can play a role in phonological mediation of imitation. In particular, we showed that duration played a role in mediating which tone category was imitated. The perception results and the baseline production results in this study showed the role of duration in distinguishing the flat and the falling tones. The role of duration in cuing the contrast is also reflected in imitation: shorter durations resulted in more falling tone imitations, as in the perception results, where a shorter duration increased falling tone responses (Blicher et al., 1990; Yang, 1989) and the category boundaries along the F0 continuum were located at roughly the same positions in the perception and imitation results (collected from different groups of participants). This correspondence between perception and imitation is consistent with the identification and imitation results of VOT in Flege and Eefting (1988), and compatible with their claim that stimuli were categorized before imitation. Nevertheless, it is also plausible that the production phase may exert a mediating effect. In the context of everyday linguistic interactions, for the production material used in this experiment (a tri-syllabic phrase with the initial one being the focus), individuals tend to articulate clear tonal patterns for the target word. This habitual practice, deeply embedded in speech communication, raises the prospect that if listeners

exhibit a nuanced perception of phonetic details for these tones (such as the perception for other segments in Kong & Edwards, 2016; Kapnoula et al., 2017; Kim et al., 2020), they may be less skilled in faithfully reproducing these subtleties during the production phase. A production-based explanation would also have to explain why the F0 cue is most strongly affected by this process. Such an explanation could perhaps come from the production distributions we observed in the baseline condition. The distribution of baseline durations was more overlapping than the baseline F0 ranges. This means that speakers had more practice with producing ambiguous durations than with producing ambiguous F0 ranges. The current study can't rule out this possibility.

Both Kim and Clayards (2019) and the present study observed that syllable duration was imitated in a more linear way. In both experiments, duration serves as a non-primary cue to the phonological contrast. One possibility is that imitation of non-primary cues may not be as influenced by the phonological effect. Alternatively, it is possible that syllable duration is a special feature that is particularly sensitive to imitation (as reviewed in Pardo (2010)) and less constrained by phonological contrasts than other features. One potential motivation for duration to be especially sensitive to imitation could be its simultaneous roles in conveying phonological, prosodic and paralinguistic information. For example, it can vary with vowel quality, prosodic prominence, or speaking rate, all in the same language. To properly test this claim, it would be necessary to gather evidence from imitation studies where duration serves as a primary cue to a contrast (e.g., to Japanese long-short vowels). If the linear imitation of duration occurs when it serves as a primary cue, it would support an account assigning special sensitivity to this feature.

We should note that the current study only examined the categoricalness of supra-segmental cues. Although previous literature has documented many non-linear imitation results, this study did not compare the non-linearity of segmental cues, such as VOT directly. It would be worthwhile to make this direct comparison in a future study to strengthen the conclusions drawn by this study.

2.5.2 Role of duration in Mandarin one contrasts

As mentioned earlier, in previous literature the duration effect on the flat-falling tonal contrast was observed inconsistently. Wang and Peng (2012) and Feng and Peng (2018) didn't find a significant duration effect on either the boundary position or the sharpness of the categorization function. In contrast, Zhu et al. (2016) found that shorter duration significantly increased the perception of the falling tone. In addition, Wang et al. (2017) found that shorter duration increases the categoricalness (steeper shift) of the flat-falling tonal perception. However, they didn't observe an effect of duration on the position of the tonal boundary. One possible reason for these variations could be the differing values of duration as well as F0 falling ranges used in these experiments. For example, Zhu et al (2016) used 100, 200, 300 ms of duration whereas Wang and Peng (2012) used 300 and 500ms. Feng and Peng (2018) investigated a maximal F0 falling range of 85 Hz whereas the maximal F0 range in Wang et al. (2017) was 50 Hz. Since the effect of duration is non-primary to the flat-falling contrast, it could be sensitive to the investigated duration values and F0 falling range values. The current study observed clear effect of duration to the tonal contrast in both perception and imitation results, possibly due to our examination of a relatively larger range of F0 falls and duration values. Additionally, duration is such as complex feature of speech that it may play multiple roles in speech perception. Apart from being an intrinsic cue for tones, it may also affect the

categorization as a feature of speech rate and cognitive resource allocation. For instance, F0 variations within a 100 ms duration may be processed differently from 400 ms due to the shorter processing resource. The specific roles that duration plays can be sensitive to the actual duration values involved. Different durations may have varying effects on perception and cognitive processes, highlighting the nuanced nature of duration as a feature in speech. The whole picture of how duration impacts the flat-falling contrast can be clearer if a wider range of F0 falls and wider range of durations are manipulated in a future study.

2.5.3 Imitation of F0

This study showed that the imitation of F0 contour is mediated by the Mandarin tonal contrast. Previous studies with intonational contours had found both linear and phonologically mediated results. How can we reconcile those results with the current study? Dilley (2010) and Michalsky (2015) investigated the imitation of F0 contour (specifically F0 range) between two English intonational pairs (e.g., between H* vs L+H*) or pragmatic pairs (Interrogativity vs Questioning), and both studies found that the F0 contour imitation was linear. On the other hand, various investigations of intonational targets have found more phonologically mediated behaviour. Several experiments manipulated the timing of the peaks or valleys of intonational contours (Dilley & Brown, 2007; Pierrehumbert & Steele, 1989; Redi, 2003; Tilsen et al., 2013). Interestingly, almost all these F0 alignment imitation results exhibited non-linearity. These results seem to indicate that at least in intonation, F0 alignment more easily induces phonological mediation in perception and production than other cues, such as F0 range. The discrepancy between the results in Dilley (2010) and Michalsky (2015) and the current result – both of which manipulated F0 range – highlights that the contrast between two Mandarin tones differs significantly from the two intonational targets or the two

pragmatic functions in the previous studies. Our results appear to support the claim in Dilley (2010) that the intonation pairs tested in Dilley (2010) and Michalsky (2015) were not phonologically categorical distinctions.

2.5.4 Gradiency and individual differences

Despite seeing mostly categorical imitation of F0 range in this study, we found that there were also phonetic details maintained in imitation. Although the last three steps of F0 range were consistently imitated as the falling tone, the different falling ranges in the stimuli were reproduced as well to some extent (as shown in Figure 2.6). The results are consistent with a recent finding of Mandarin tones – Qin et al. (2019) found that native speakers can perceive and make use of the within-category F0 variation in lexical activation. Furthermore, the pattern is also compatible with recent findings of sub-phonemic imitations. Zellou and colleagues (Zellou, Scarborough & Nielsen, 2016; Zellou, Dahan & Embic, 2017) observed that participants were sensitive to degree of coarticulatory nasality which is not contrastive in English. Furthermore, Nielsen (2011) found that the lengthened VOT on /p/, which is also a phonetic rather than phonological variation, was imitated by participants. However, there is less within-category imitation in the flat tone category, indicating asymmetries for certain ranges of cues for which imitation is more phonetic.

This result also highlights that we observed some gradiency even within the cue that was imitated most categorically. While many early studies of speech perception tended to emphasize the categorical nature of speech perception (e.g. Liberman et al., 1957; Repp, 1984), there have been many studies that have emphasized the gradient nature of perception as well (see McMurray, 2022 for a review). For example, Kong and Edwards (2016) and Kapnoula

et al. (2017) observed that when presented with a continuous scale, rather than two categories, many listeners responded in a gradient way to the continuum of stimuli in a typical categorical perception study. Interestingly, they also noted that some listeners were more gradient than others.

We also found that some of our participants imitated the F0 targets categorically, some were good at imitating each F0 step linearly, and most of them were using a hybrid method: the imitation distinguished the tonal categories and also kept the fine-grained within-category differences. Although it is also possible that some individual-level differences in our study are due to the participants' varying accents, these results are in line with previous studies that found listeners varied in how gradient or categorically they behaved in perception tasks (e.g. Kong & Edwards, 2016; Kapnoula et al., 2017; Kim et al., 2020). Individual differences in imitation have also been observed in Kim and Clayards (2019). Future studies are needed to address whether individual differences in the gradiency of phonetic imitation is driven by gradiency in perception.

The stimulus in this study contained not only the target continuum, but also a carrier to avoid phrase-final creak. The presence of the carrier may influence the perceptual and/or imitative gradiency. Additionally, while we chose a neutral tone to connect with the target syllable to reduce the tonal co-articulation effects, there might be a delay in realizing the F0 target in production. It will be interesting to examine the effect of carrier on the categoricalness of imitation in future studies.

2.5.5 Conclusion

In summary, through an identification and an imitation experiment of Mandarin flat-falling tonal continua varying continuously in two dimensions, this study revealed that in phonetic imitation cues that are often described as 'supra-segmental' are not necessarily imitated accurately. We found that the imitation of F0 range was mostly categorical while the imitation of duration was more gradient (linearly following the stimuli). Since both cues are supra-segmental, the different imitation patterns observed for F0 range and duration may be attributed to their roles as primary and non-primary cues, respectively, to the tonal contrast. Alternatively, there may be something special about syllable or vowel duration that lends itself well to gradient imitation. The correspondence between the identification and imitation results is compatible with previous findings that stimuli are categorized before being imitated. We also observed that individuals show variability in how categorically they imitate the F0 range, and that most participants did exhibit some gradiency for different F0 falling ranges within the falling category.

References

Alivuotila, L., Hakokari, J., Savela, J., Happonen, R.-P., & Aaltonen, O. (2007). Perception and imitation of Finnish open vowels among children, naïve adults, and trained phoneticians. *Proceedings of ICPhS XVI*, 361–364.

Babel, M., Delaney, M., & Savji, S. (2011). Implicit and explicit phonetic imitations in single-word shadowing. *The Journal of the Acoustical Society of America*, *129*(4), 2657–2657. https://doi.org/10.1121/1.3588875

Babel, M. E. (2009). *Phonetic and social selectivity in speech accommodation*. University of California, Berkeley.

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice*. Cengage Learning.

Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics*, *18*(1), 37–49. https://doi.org/10.1016/S0095-4470(19)30357-2

Boersma, P. (2006). Praat: Doing phonetics by computer. *Http://Www. Praat. Org/*.

Bürkner, P.C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Chao, Y. R. (1965). *A grammar of spoken Chinese*. Univ of California Press.

Chistovich, L., Fant, G., de Serpa-Leitao, A., & Tjernlund, P. (1966). Mimicking of synthetic vowels. *Quarterly Progress and Status Report, Speech Transmission Lab, Royal Institute of Technology, Stockholm*, *1*(2), 1–18.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108(3), 804–809. https://doi.org/10.1016/j.cognition.2008.04.004

Cohen, J. (1988). Set correlation and contingency tables. *Applied Psychological Measurement*, *12*(4), 425–434.

Coupland, J., Coupland, N., & Giles, H. (1991). Accommodation theory. Communication, context and consequences. *Contexts of Accommodation*, 1–68.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

De Looze, C., & Hirst, D. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. *Speech Prosody*, 4.

Denes, P. (1955). Effect of duration on the perception of voicing. *The Journal of the Acoustical Society of America*, *27*(4), 761–764.

Dilley, L. C. (2010). Pitch Range Variation in English Tonal Contrasts: Continuous or Categorical? *Phonetica*, *67*(1–2), 63–81. https://doi.org/10.1159/000319379

Dilley, L. C., & Brown, M. (2007). Effects of pitch range variation on f0 extrema in an imitation task. *Journal of Phonetics*, *35*(4), 523–551. https://doi.org/10.1016/j.wocn.2007.01.003

D'Imperio, M., & Sneed, J. (2015). Phonetic detail and the role of exposure in dialect imitation. *18 Th International Congress of Phonetic Sciences*.

Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology*, 4, 346. doi: 10.3389/fpsyg.2013.00346

Elert, C. C. (1964). Phonologic studies of Swedish quantity. *Uppsala: Almqvist & Wiksell*.

Feng, Y., & Peng, G. (2018). The effect of duration on categorical perception of Mandarin tone and voice onset time. *TAL2018, Sixth International Symposium on Tonal Aspects of Languages*, 164–168. https://doi.org/10.21437/TAL.2018-33

Flege, J. E., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. *The Journal of the Acoustical Society of America*, *83*(2), 729–740. https://doi.org/10.1121/1.396115

Garnier, M., Lamalle, L., & Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00600

Giles, H., Coupland, N., & Coupland, I. (1991). 1. Accommodation theory: Communication, context, and. *Contexts of Accommodation: Developments in Applied Sociolinguistics*, *1*.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251. https://doi.org/10.1037/0033-295X.105.2.251

Hao, Y.-C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics*, *54*, 151–168. https://doi.org/10.1016/j.wocn.2015.10.003

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013–3022. https://doi.org/10.1121/1.1323463

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. The Journal of the Acoustical Society of America, 119(5), 3059–3071. https://doi-org.proxy3.library.mcgill.ca/10.1121/1.2188377

Jongman, A., Wang, Y., Moore, C. B., & Sereno, J. A. (2006). Perception and production of Mandarin Chinese tones. In P. Li, L. H. Tan, E. Bates, & O. J. L. Tzeng (Eds.), *The Handbook of East Asian Psycholinguistics* (pp. 209–217). Cambridge University Press. https://doi.org/10.1017/CBO9780511550751.020

Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., & McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(9), 1594–1611. https://doi.org/10.1037/xhp0000410

Kent, R. D. (1973). The imitation of synthetic vowels and some implications for speech memory. *Phonetica*, *28*(1), 1–25. https://doi.org/10.1159/000259442

Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, *34*(6), 769–786. https://doi.org/10.1080/23273798.2019.1582787

Kim, D., Clayards, M., & Kong, E. J. (2020). Individual differences in perceptual adaptation to unfamiliar phonetic categories. *Journal of Phonetics*, *81*, 100984. https://doi.org/10.1016/j.wocn.2020.100984

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57. https://doi.org/10.1016/j.wocn.2016.08.006

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), 10. https://doi.org/10.5334/labphon.83

Leung, K. K. W., & Wang, Y. (2020). Production-perception relationship of Mandarin tones as revealed by critical perceptual cues. *The Journal of the Acoustical Society of America*, *147*(4), EL301–EL306. https://doi.org/10.1121/10.0000963

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358. doi: 10.1037/h0044417

Lin, Y., Yao, Y., & Luo, J. (2021). Phonetic accommodation of tone: Reversing a tone merger-in-progress via imitation. *Journal of Phonetics*, *87*, 101060. https://doi.org/10.1016/j.wocn.2021.101060

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech*, *2017*, 498–502.

MacLeod, B., & Di Lonardo Burr, S. M. (2022). Phonetic imitation of the acoustic realization of stress in Spanish: Production and perception. Journal of Phonetics, 92, 101139. https://doi.org/10.1016/j.wocn.2022.101139

McMurray, B. (2022). The myth of categorical perception. *The Journal of the Acoustical Society of America, 152(6), 3819-3842.* doi: 10.1121/10.0016614

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental*

Psychology: Human Perception and Performance, 34(6), 1609-1631. https://doi.org/10.1037/a0011747

Michalsky, J. (2015). Pitch scaling as a perceptual cue for questions in German. *Sixteenth Annual Conference of the International Speech Communication Association*.

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. https://doi.org/10.1016/j.cognition.2008.08.002

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142. https://doi.org/10.1016/j.wocn.2010.12.007

Nye, P. W., & Fowler, C. A. (2003). Shadowing latency and imitation: The effect of familiarity with the phonetic patterning of English. *Journal of Phonetics*, *31*(1), 63–79. https://doi.org/10.1016/S0095-4470(02)00072-4

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393. https://doi-org /10.1121/1.2178720

Pardo, J. S. (2010). *Expressing Oneself in Conversational Interaction*. 14.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637–659. https://doi.org/10.3758/s13414-016-1226-0

Pierrehumbert, J. B., & Steele, S. A. (1989). Categories of tonal alignment in English. *Phonetica*, *46*(4), 181–196. DOI: 10.1159/000261842

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, *13*(2), 253–260. https://doi.org/10.3758/BF03214136

Postma-Nilsenová, M., & Postma, E. (2013). Auditory perception bias in speech imitation. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00826

Qin, Z., Tremblay, A., & Zhang, J. (2019). Influence of within-category tonal information in the recognition of Mandarin-Chinese words by native and non-native listeners: An eye-tracking study. *Journal of Phonetics*, *73*, 144–157. https://doi.org/10.1016/j.wocn.2019.01.002

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria.

Redi, L. (2003). Categorical effects in production of pitch contours in English. *Proceedings of the 15th International Congress of the Phonetic Sciences*, 2921–2924.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In *Speech and language* (Vol. 10, pp. 243–335). Elsevier. https://doi.org/10.1016/B978-0-12-608610-2.50012-1

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L., & Nguyen, N. (2013). Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00422

Schertz, J., Adil, F., & Kravchuk, A. (2023). Underpinnings of explicit phonetic imitation: Perception, production, and variability. Glossa Psycholinguistics, 2(1). https://doi.org/10.5070/G601123

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. WIREs Cognitive Science, 11(2). https://doi.org/10.1002/wcs.1521

Tilsen, S., Burgess, D., & Lantz, E. (2013). Imitation of intonational gestures: A preliminary report. *Cornell Work. Pap. Phon. Phonol*, 1–17.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cognitive Science, 34(3), 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Wagner, M. (2021). *Prosody lab Experimenter* [JavaScript]. Prosodylab. https://github.com/prosodylab/prosodylabExperimenter (Original work published 2020)

Wagner, M. A., Broersma, M., McQueen, J. M., Dhaene, S., & Lemhöfer, K. (2021). Phonetic convergence to non-native speech: Acoustic and perceptual evidence. *Journal of Phonetics*, *88*, 101076. https://doi.org/10.1016/j.wocn.2021.101076

Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology*, *6*, 546. https://doi.org/10.3389/fpsyg.2015.00546

Wang, D., & Peng, G. (2012). *Effects of pitch range and duration on tone categorical perception*. https://www.isca-speech.org/archive/tal_2012/tl12_O2-03.html

Wang, T., Liu, J., Lee, Y., & Lee, Y. (2020). The interaction between tone and prosodic focus in Mandarin Chinese. *Language and Linguistics. 語言暨語言學*, *21*(2), 331–350. https://doi.org/10.1075/lali.00063.wan

Wang, Y., Yang, X., & Liu, C. (2017). Categorical Perception of Mandarin Chinese Tones 1–2 and Tones 1–4: Effects of Aging and Signal Duration. *Journal of Speech, Language, and Hearing Research*, *60*(12), 3667–3677. https://doi.org/10.1044/2017_JSLHR-H-17-0061

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, *93*(4), 2152–2159. https://doi.org/10.1121/1.406678

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*(3–4), 220–251. https://doi.org/10.1016/j.specom.2005.02.014

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *The Journal of the Acoustical Society of America*, *111*(3), 1399–1413. https://doi.org/10.1121/1.1445789

Yang, Y. (1989). The vowels and the perception of Chinese tones. *ACTA Psychologica Sinica* (1): 29-33. (In Chinese)

Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic Imitation from an Individual-Difference Perspective: Subjective Attitude, Personality and "Autistic" Traits. *PLoS ONE*, *8*(9), e74746. https://doi.org/10.1371/journal.pone.0074746

Zellou, G., Dahan, D., & Embick, D. (2017). Imitation of coarticulatory vowel nasality across words and time. *Language, Cognition and Neuroscience*, 32(6), 776–791. https://doi.org/10.1080/23273798.2016.1275710

Zellou, G., Scarborough, R., & Nielsen, K. (2016). Phonetic imitation of coarticulatory vowel nasalization. *The Journal of the Acoustical Society of America*, *140*(5), 3560–3575. https://doi.org/10.1121/1.4966232

Zetterholm, E. (2007). Detection of Speaker Characteristics Using Voice Imitation. In *Speaker Classification II* (Vol. 4441, pp. 243–257). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74122-0_19

Zhu, W., Wei Y., Wu L. & Wang J. (2016). The Effect of Pitch Range and Tone Duration on Chinese Tone Perception by L2 Learners.*Chinese Language Learning* (2): 83-92 (In Chinese)

Appendix 2.A Priors for the Bayesian models.

Priors of the mixture model for F0 range imitation

```
prior  = c(
        prior(normal(0, 1), Intercept, dpar = mu1),
        prior(normal(4, 1), Intercept, dpar = mu2),
        prior(cauchy(0, 1), sd, dpar = mu1),
        prior(cauchy(0, 1), sd, dpar = mu2)
```

```
    )
```

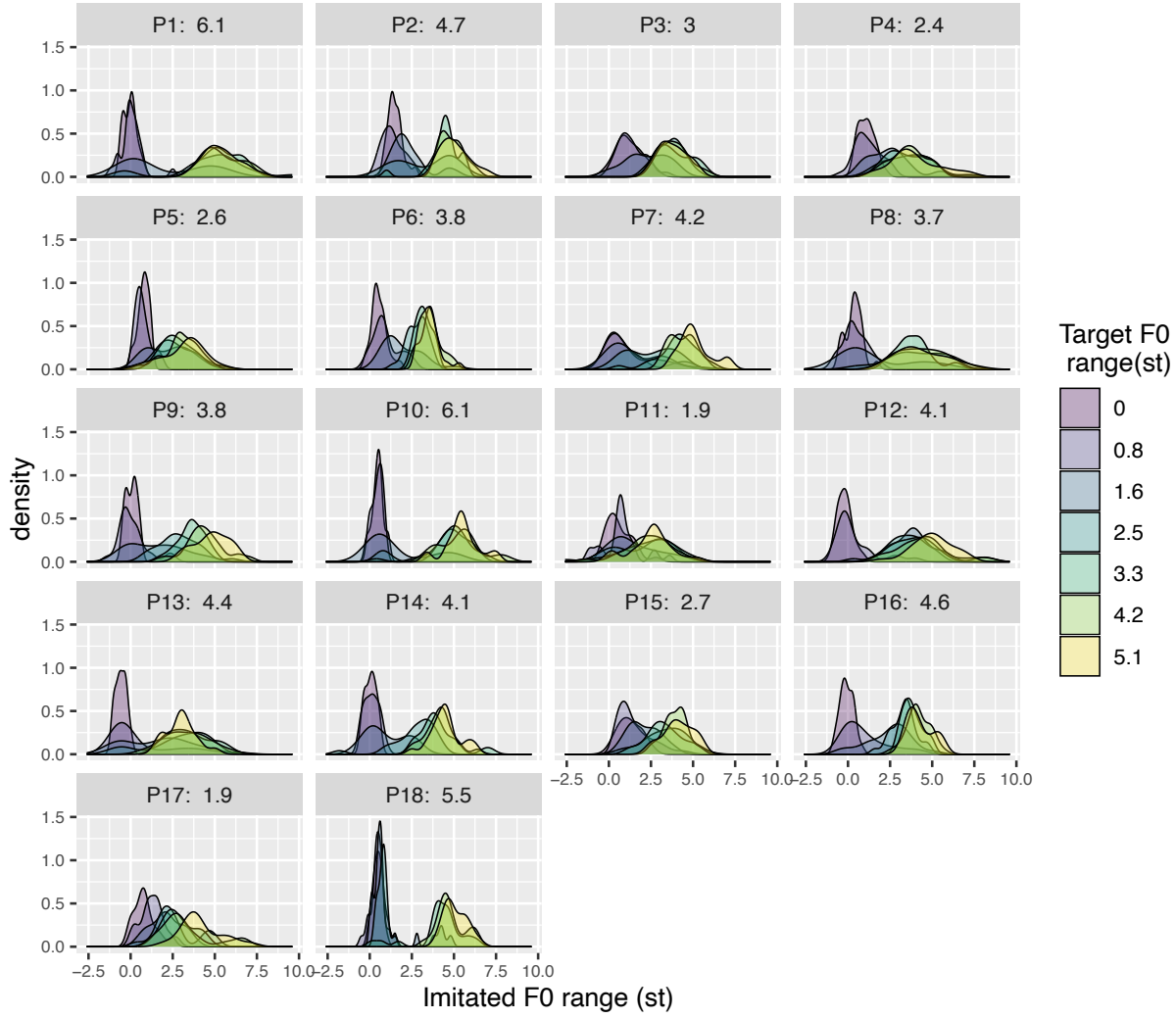
Priors of the simple Gaussian model for F0 range imitation

```
    prior = c(
        prior(normal(2.5, 2), class=Intercept),
        prior(normal(0, 1.5), class=b, dpar = sigma, coef = f0_range_target)
    )
```

Priors of the mixture model for duration imitation

```
    prior = c(
        prior(normal(0.2, 0.1), Intercept, dpar = mu1),
        prior(normal(0.26, 0.2), Intercept, dpar = mu2)
    )
```

Appendix 2.B Distribution of F0 range imitation with Cohen's d values for 18 participants



Preface to Chapter 3

This dissertation explores the imitation mechanism of F0 and duration, both of which are 'supra-segmental' dimensions closely tied to 'supra-segmental' events. The study in Chapter 2 examined the imitation of F0 and duration under the effect of the Mandarin flat-falling tonal contrast. The results do not support the 'supra-segmental priority' hypothesis. Through the Mandarin flat-falling tonal continua, Chapter 2 revealed that the imitation of F0 is mediated by phonological contrast: F0 fall incrementally increases in the stimuli however the imitations have a bimodal distribution.

As for the other dimension which is closely tied with 'supra-segmental' events like F0, Chapter 2 found that, unlike F0, the imitation of duration exhibited a linear distribution for the incrementally increased durations in the stimuli. This result is consistent with the findings of Kim and Clayards (2019), where vowel continua were utilized in the imitation experiment. Results in the two studies seem to suggest that duration imitation may not be mediated by the phonological contrast. However, it's important to note that in both studies, duration only plays a weak role in distinguishing the examined phonological contrast. The impact of phonological effects on a cue may depend heavily on the cue's importance to the phonological contrast (Dupoux et al., 1999). Therefore, it remains unclear whether the linear imitation pattern of duration stems from the unique characteristics of duration itself or if it results from the relatively minor role of duration in those phonological contrasts. To investigate if the observed linear imitation of duration thus far is limited by its non-primary role it plays in the contrasts, and to gain a comprehensive understanding of the imitation mechanism of

'suprasegmental' dimensions, an imitation study on a contrast where duration plays an important role is necessary.

Chapter 3 is designed to examine the mechanism of duration imitation by extending its role to a strong cue to the phonological contrast. To facilitate a more parallel comparison with F0 imitation, as explored in Chapter 2, we continue to use tonal contrasts. Taiwanese Southern Min (TSM) is a language containing 7 lexical tones (Norman 1988). Two of them are checked tones, which contain a final unreleased stop and are shorter, while the other five are unchecked tones, which do not contain final stops and are longer. Recent studies have suggested that duration, as a cue distinguishing between checked and unchecked tones, is possibly playing an increasing role (Pan, 2017). This suggests that the checked-unchecked tone contrast could be a material for further examination of duration imitation properties.

The study in Chapter 3 consists of two experiments. In the first experiment, the role of duration in distinguishing the checked-unchecked tones was established. Results showed that duration plays a key role in the mid register checked-unchecked tonal contrast (T3-T33), but not in the high register checked-unchecked tonal contrast (T5-T55). As such, in the second experiment, we utilized the T3-T33 contrast to examine duration imitation, through a similar method as in Chapter 2. Given the potential variations in the result of duration imitation, the following predictions can be made,

- Prediction A

  If the imitation of duration along the T3-T33 contrast in TSM remains linear, we may conclude that duration is a unique cue, capable of bypassing phonological mediation in phonetic imitation.

- Prediction B

  If the imitation of duration along the T3-T33 contrast is no longer linear but shows a bimodal pattern, it suggests that, similar to F0 and VOT, duration is subject to phonological mediation.

When understanding the role that phonology plays in the imitation of 'suprasegmental' dimensions, linking the results of F0 imitation and duration imitation, Prediction A pointed towards a loose association of dimension type. Although F0 and duration are both labeled as 'suprasegmental' dimensions, they demonstrate different performance in resisting the phonological mediation in imitation. Prediction B, however, suggests a strong influence of phonological mediation, consistent with observations from the imitation of any dimension examined in the literature so far.

Chapter 3

Study 2: The role of syllable duration in the perception and imitation of checked vs unchecked tones in Taiwanese Southern Min

3.1 Introduction

In the past decades, substantial work has been done to explore how listeners perceive speech segments. Among others, the contrast of voiced vs. voiceless stops, which is cued primarily by voice onset time (VOT), vowel quality contrasts, which are distinguished primarily by formants, and full lexical tone contrasts, which are distinguished primarily by fundamental frequencies (F0) have received ample investigations. The present study investigates the perception and imitation of a checked vs. full tonal contrast that is not primarily cued by F0, which is examined to a lesser extent in the previous literature.

Taiwanese Southern Min (TSM) contains 2 checked tones and 5 full tones (i.e., unchecked tones). The checked tones are distinguished from unchecked ones by having a final unreleased stop and shorter syllable duration (Norman, 1988). Recent studies found that the importance of the final unreleased stop seems to be decreasing over time (Pan, 2017), which accordingly suggests that the role of syllable duration in distinguishing checked vs unchecked tones may be increasing. The current study has two goals. First, to examine how important duration is in the checked-unchecked tonal contrasts in TSM.

Secondly, the present study aimed to explore how native speakers behave in the phonetic imitation of the duration cue, to the checked and unchecked tonal contrast. Previous studies found that in full tone contrasts, such as the flat vs. falling tonal contrast in Mandarin, the main cue F0 was imitated non-linearly (Zhang et al., 2023). This means that when F0 range varies evenly, the imitation of these tones is uneven due to the influence of tonal phonology. However, less is known about how syllable/vowel duration is imitated when it serves as an important cue to a phonological contrast. Hence, the goal of this study was to further examine the phonological effect in syllable duration imitation across two phonological categories contrasted by syllable duration.

3.1.1 Phonetic imitation of syllable duration

Phonetic imitation in general means the phenomenon that a speaker adjusts their speech production to the interlocutor, resulting in an increased similarity between the speakers (see Pardo et al. (2017) for a review). In recent years, it has been used as a paradigm to test the phonological nature of segmental or supra-segmental contrasts (Flege & Eefting, 1988; Kim & Clayards, 2019; Kwon, 2019; Mitterer & Ernestus, 2008; Nielsen, 2011). For example, Flege and Eefting (1988) investigated how a VOT continuum between -60 ms to 90 ms was imitated by three groups of speakers: monolingual English speakers, monolingual Spanish speakers and Spanish-English bilinguals. Results showed that all three groups of participants imitated the steps of the continuum in a categorical way – all groups imitated the VOTs into two or three modes or clusters, and the VOT steps between the modes were imitated with larger variations which suggest less precise imitations. The number and the center values of the modes reflected each group's phonetic categories along the VOT dimension: either pre-voiced, voiceless unaspirated, or voiceless aspirated. Such a non-linear or categorical

imitation pattern of an acoustic continuum is thought to show the phonological mediation effect on the dimensions varied in the continuum (Chistovich et al., 1966; Flege & Eefting, 1988; Kent, 1973; Kim & Clayards, 2019; Repp & Williams, 1985; Zhang et al., 2023, 2024).

In a recent study, Zhang et al. (2023) investigated the phonological effect on the imitation of F0 and syllable duration in the Mandarin flat vs. falling tonal contrast. F0 was the most important cue and duration was a less important cue to the contrast. As was found in Flege and Eefting (1988), results showed that the important cue, F0, was imitated by native speakers in a categorical manner. However, the less important cue, syllable duration, was imitated by the speakers in a linear manner. Specifically, each step of the duration continuum was imitated differently, without any clustering into modes or categories. The linear imitation of a syllable duration continuum was also observed in Kim and Clayards (2019) who investigated the imitation of /ɛ/ and /æ/ in English. These results seem to suggest very little mediation of syllable duration by the phonological effect. In another study, Podlipskỳ and Simácková (2015) examined the effect of phonological contrast on imitation through the Czech long vowel /u:/ which contrasts phonologically with short /u/. Podlipskỳ and Simácková (2015) created both shortened and lengthened versions of /u:/ and used them in a shadowing task. They found that both the shortened and extended /u:/ were imitated, which aligns with the duration imitation results in Zhang et al. (2023) and Kim and Clayards (2019). However, in a follow up study where a subset of the stimuli from Podlipskỳ and Simácková (2015) were examined within a carrier phrase, Kovaříková (2023) failed to observe imitation of the shortened /u:/, only noting imitation of the lengthened /u:/. Therefore, the phonological effect on duration imitation remains uncertain.

In other studies which focused on non-phonological factors affecting imitation, such as social effects, it was consistently found that listeners imitated or converged to durational variations of vowels, full tones, lexical stress, and sentences (Coupland et al., 1991; Delvaux & Soquet, 2007; MacLeod & Di Lonardo Burr, 2022; Pardo, 2010; Street & Cappella, 1989; Wagner et al., 2021; Zetterholm, 2007). This leads to the question of whether the accurate/linear imitation of duration that has been observed so far is due to a specific property of duration, which would make it behave differently from any of the cues that have been investigated, such as F0 or VOT. Importantly, duration in both Zhang et al. (2023) and Kim and Clayards (2019) was a weak cue to the contrast investigated. In contrast, vowel duration in Podlipský and Simácková (2015) a key cue for the vowel contrast. However, they only examined 2 modifications of the long vowel, without testing the short /u/ category. In other words, none of the previous studies have investigated duration imitation for a contrast where duration plays an important role in the phonological contrast, as F0 does in the Mandarin flat-falling tonal contrast. If, for a contrast that is mainly cued by duration, the imitation of duration along the contrast is still linear, we may conclude that duration is a very special cue that can bypass the phonological mediation in phonetic imitation. If, on the contrary, when duration plays an important role to a phonological contrast, the imitation of it is no longer linear but shows a categorical pattern, it suggests that similar to F0 and VOT, duration imitation is modulated by the phonological effect. Moreover, as indicated in Zhang et al. (2024), there might be another possibility that it can show both linear and categorical characteristics, which indicates that duration imitation is mediated by phonological categories while also retaining within-category phonetic details.

To the best of our knowledge, no one has examined the imitation of a vowel/syllable duration continuum in a case where it plays an important role in a phonological contrast. The present study aims to fill this gap and thus provide more understanding of the imitation mechanism.

3.1.2 Checked vs unchecked tones in Taiwanese Southern Min

To examine our question, we utilized the contrast of checked vs. unchecked tones in Taiwanese Southern Min (TSM). TSM, also known as Taiwanese Hokkien, is the second most spoken language in Taiwan (Ang, 2013). However, due to the widespread use of Mandarin, TSM use is declining, particularly among the younger generation (Chen, 2010; Ding, 2016). TSM has a rich tonal system including seven lexical tones, two of which are checked tones occurring in CV [p, t, k, ʔ] syllables where the final stops are unreleased (Norman, 1988). Unchecked tones, on the other hand, occur in CV(N) syllables. The full tonal space is described in Pan (2017) using numbers to indicate pitch height following Chao (1930). Pan (2017) also shows that the checked tones (T5, T3) are shorter than the unchecked tones (T55, T13, T51, T31, T33).

In recent years, there has been growing attention to the sandhi schemes of TSM (Chien & Jongman, 2019; Kuo, 2013; Pan, 2017), and the acoustic properties of checked and unchecked tones respectively (Pan, 2005; Pan et al., 2011, 2016). However, the investigations on how listeners perceive the checked vs. unchecked tone pairs in TSM has been limited. There are two pairs of contrasting checked and unchecked tones for mid and high registers, respectively: T3 vs T33 and T5 vs T55. Phonologically, the distinction between checked and

unchecked tones are the final unreleased stop. However, acoustically, the duration difference is also very salient: unchecked tones are about twice as long as checked tones.

Pan (2017) reported cases of coda deletion, in which over 80% of /ʔ/ codas were deleted in checked tones especially for younger talkers. This indicates a possibly increasing role for duration in checked tone perception. Furthermore, previous studies using electroglottographic and acoustic data, have shown that, while final stop closures may be absent, energy dipping and irregular glottal vibration can still be observed on the preceding vowels (Pan et al., 2016), suggesting that the irregular spectral vibration (i.e., glottalization) can be a cue to recognize the checked tone. Thus, checked tones in TSM which end with voiceless stops and are shorter and often accompanied by glottalization are similar to voiceless stop codas in English which are often cued by a shorter preceding vowel and the presence of glottalization (Penney et al., 2018). In this study, the extent to which listeners make use of duration and glottalization in perceiving the TSM checked vs. unchecked tones will be investigated, for both registers.

There are relatively fewer investigations of the perception than the production of checked tones or checked syllables in the previous literature (see Chai, 2022, for a review). From the few investigations of checked tone perception in other languages or Chinese dialects, duration, voice quality (such as glottalization) and F0 are the most important cues reported. As one can imagine, the relative importance of the three cues varies across languages: In Burmese, voice quality and duration might be the cues distinguishing the checked tone from the rest of the tones (Gruber, 2011); for both Southern and Northern Vietnamese, glottalization was more important than F0 (Brunelle, 2009); for White Hmong, F0 and duration were more important than glottalization (Garellek et al., 2013); perhaps the most

relevant comparison is Xiapu Min, in which duration seems to be the most important (Chai, 2021), followed by F0, with glottalization being the least important (Chai 2022).

In terms of TSM, both the checked and unchecked tones have a falling contour acoustically. Additionally, Lin (2022) showed that when the durations are normalized, the checked tones and unchecked tones had very similar F0 contours, suggesting that F0 may not be an important cue to the checked vs. unchecked tonal contrast. As we are interested in the role of contrast in the imitation of duration, this study first investigates how important the duration cue is in perceiving the checked vs. unchecked tonal contrast in TSM, relative to glottalization. If duration strongly modulates perception of the checked vs. unchecked tones, it provides us a chance to examine the question of whether duration imitation is also strongly modulated by the phonological contrast.

3.1.3 The present study

Two experiments were conducted to investigate the role of syllable duration in the categorization and imitation of checked vs. unchecked tones in TSM.

The first experiment is a perception experiment for the checked vs. unchecked tonal contrast of mid (i.e., T3 vs T33) and high (i.e., T5 vs T55) registers. We selected the syllable structure of CV[ʔ] for the checked tones as Pan (2017) observed that the final coda [ʔ] was deleted more frequently than [p, t, k] in TSM checked tones. This suggests that duration plays a relatively larger role in checked tones with coda [ʔ] than with the other three codas. As detailed below, we created syllable duration continua between the checked tone and the unchecked tone on the glottalized vowel base and the non-glottalized vowel base, respectively.

Categorization of the duration continua in the two base vowels will be compared so that the role of duration can be evaluated relative to the role of glottalization.

If duration turns out to be an important cue for listeners to distinguish between the checked vs unchecked tones, then the continuum will be used in Experiment 2, which is an imitation experiment, to examine the linearity of the duration imitation (i.e., the phonological effect in imitation) across checked vs unchecked tone continua.

3.2 Experiment 1: perception

3.2.1 Method

3.2.1.1 Materials

The segmental context was controlled for both checked and unchecked tones for both registers. Specifically, the syllable /la/ was used for T33 ('扒', meaning 'stir') and T55 ('拉', meaning 'pull') while /laʔ/ was used for T3 ('垃', meaning 'garbage') and T5 ('蠟', meaning 'wax'). In the literature, /l/ in TSM was frequently described as a voiced stop with either a central or lateral release, so it is sometimes transcribed as [ɾ] or [l] (Zhang, 1989). A female speaker of TSM recorded each of the four syllables naturally, in a sound-attenuated booth. From these recordings, one sample of each tone was selected with the least perturbation and the smoothest amplitude envelope to generate the checked-unchecked tonal continua.

To evaluate how important duration is to the checked tone perception, we created two types of tonal continua. The first type (All-cue Continuum) was designed to determine if participants could clearly distinguish between checked and unchecked tones, under ideal

conditions. The All-cue Continuum was created by morphing between the checked and unchecked tone endpoints, hence every contributing cue (both the spectrum and syllable duration) should be proportionally manipulated. Tandem-Straight (Kawahara, 2006), a speech analysis, modification and resynthesis framework, was used to create the All-cue Continuum. The spectrograms of the endpoint tokens are shown in Figure 3.1, where the syllable onset and offset are marked by the vertical box edges.

The role of syllable duration in the checked vs. unchecked contrasts was examined more directly by the second type (Duration Continuum). Duration Continua were created by manipulating only the syllable duration of each of the checked and unchecked tones. Using the Pitch Synchronous Overlap and Add (PSOLA) algorithm in Praat (Boersma, 2020), we generated Duration Continua by (1) increasing the duration of the checked tone token or by (2) decreasing the duration of the unchecked tone token. The main difference between All-cue Continua and Duration Continua is whether the two cues (vowel glottalization and duration) are varied orthogonally (Duration Continua) or concomitantly (All-cue Continua) in the stimuli.

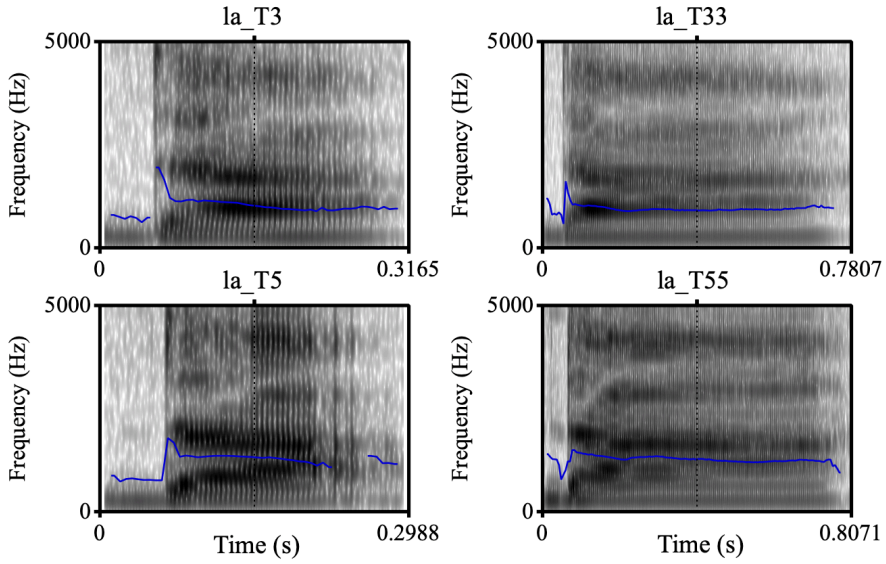


Figure 3.1: Spectrograms of base tokens (top left: T3; top right: T33; bottom left: T5; bottom right: T55).

Altogether there were six continua used in this experiment: Duration Continuum based on T3, Duration Continuum based on T33, All-cue Continuum of the mid register (i.e., register 3), and the same three continua for the high register (i.e., register 5).

There were seven steps in each duration continuum. Endpoint durations were 160 ms (the checked tone) and 340 ms (the unchecked tone), with the step size being around 30 ms. The duration values were selected starting from previous production results of TSM tones (unpublished data from a collaborator on the current study) and adjusted after pilot testing to ensure that the ambiguous points in the continua were close to the middle continuum step.

As all of the participants were expected to be proficient in Mandarin but with varying levels of proficiency in TSM, we created a TSM proficiency questionnaire that included questions such as age of acquisition, frequency of use, circumstances of use, listening level and speaking level (assessed between 1-7).

All stimuli, data and the scripts used to analyze the data can be found online, currently hosted on the OSF at https://osf.io/f54g6/?view_only=f20ae6efc0ba4d03afafd6a1af5f94f2.

3.2.1.2 Participants

We recruited 17 TSM speakers (10M, 7F, mean age = 23, sd = 1.6) at the Experimental Phonology Lab, National Yang Ming Chiao Tung University for the experiment. Four participants were excluded because their self-reported TSM listening level (from the pre-task questionnaire) was 3 or lower, leaving 13 speakers' data for analysis. None of them reported any hearing-related issues.

3.2.1.3 Procedure

A two-alternative forced choice (2AFC) paradigm was used. All stimuli were randomized together in a single block of trials. Since there were twice as many stimuli of the Duration Continuum type as the All-cue Continuum type, we repeated the stimuli in the All-cue Continuum within the block. Hence there were 8 continua in each block and the block repeated four times. Participants were allowed to take a break after each block. Altogether there were 224 (8 [continua]*7 [steps]*4 [blocks]) trials in the experiment. Before the main task, the participants completed a questionnaire related to their use of TSM. Before the formal experiment, participants were shown instructions with audio examples and completed 6 practice trials.

3.2.2 Statistical analysis

Statistical analyses of the identification results are carried out with Bayesian mixed effects logistic regression models, implemented using *brms* (Bürkner, 2018), an R-based front-end to the Stan programming language (Stan Development Team, 2019, 2021), in R (R Core Team, 2013). The dependent variable was the tone type response (checked T3 and T5 were mapped to 0 and unchecked T33 and T55 were mapped to 1). The specific predictors, numbers of chains and sample sizes, and prior settings are introduced in the following section. R-hat, bulk and tail ESS values of all effects in the Bayesian models were inspected to confirm that the model shows adequate sampling and convergence. In reporting results of Bayesian models in this study, the median of the posterior distribution of the estimate, the 95% credible intervals (CI), and the probability of direction (pd, computed using the R package *bayestestR*, Makowski

et al., 2019) will be given. A pd value of 95% or higher is taken to be the evidence of a credible effect.

3.2.3 Results

The categorization results for the All-cue continuum are presented in Figure 3.2. The identification rates of the extreme steps (step 1 and step 7) approach 0 or100%, forming an 'S' shape curve, although the rate of change in the mid register looks less steep than that of the high register. A Bayesian mixed effects logistic regression model was fit to the data. Predictors were *step*, *register* and the two-way interaction between them. The predictor *duration* was centered and scaled before model fitting using the *rescale()* function (Gelman, 2008). The predictor *register* was contrast coded (the mid register is coded as -0.5 and the high register is coded as 0.5). To account for individual-level variations, correlated by-participant random intercepts and random slopes of *step*, *register* and their interaction were used. The regression model was fit to data by drawing 4,000 samples in each of four Markov chains from the posterior distribution over model parameters, with a warm-up period of 1,000 samples per chain. The model was fit using weakly informative priors for fixed-effect terms: a normal distribution with a mean of 0 and a standard deviation of 10, and default priors for other parameters: the default half Student's t-distribution with 3 degrees and a brms-default scale of 2.5 for standard deviations of random intercept and random slopes, and the default LKJ prior with $\eta=1$ for correlations of random effects, in order to give lower prior probability to perfect correlations (Vasishth et al., 2018).

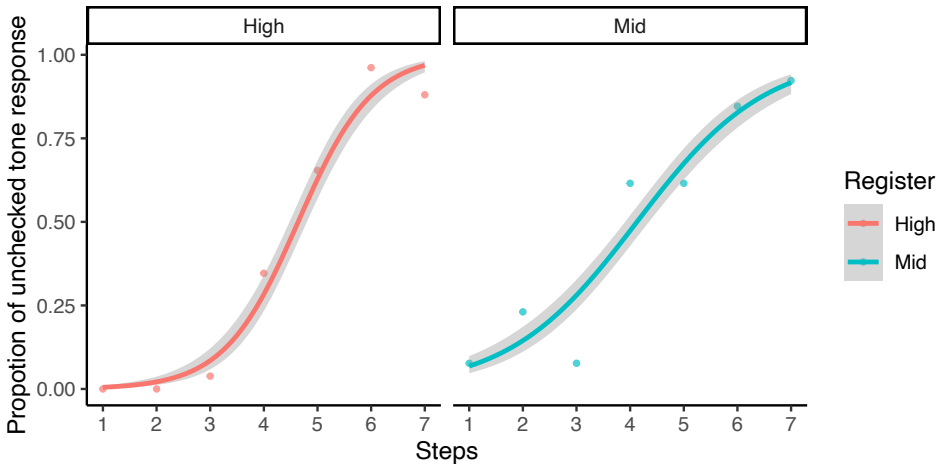


Figure 3.2: Empirical results of tone identification of the All-cue Continua, grouped by tone register. Each dot represents the average response rate across participants given the step. Each curve is a smooth from a generalized linear model to the empirical data for visualization. Shaded areas are the 95% confidence intervals of the smooth. Settings are the same for Figure 3.4 and Figure 3.5.

Results showed that larger *step* index (β = 12.51, 95% CI = [8.48, 17.24]) and lower *register* (β = -1.57, 95% CI = [-3.58, 0.26]) credibly increase the unchecked tone response. Although there is a tendency that *register* interacts with *step*, its effect is not credible (β = 5.39, 95% CI = [-2.27, 13.25]). These results (including Figure 3.2) ensure that participants are, as expected, able to distinguish the checked vs. unchecked tones in TSM, when duration and glottalization co-vary in the stimuli.

Table 3.1: Results of the mixed-effect logistic Bayesian model for the categorization of All-cue continuum. Probabilities of direction (*pd*s) of credible effects are bolded. Abbreviation: CI: confidence Interval. Same in Table 3.2-3.4.

	β	l-95% CI	u-95% CI	pd
Intercept	-0.53	-1.82	0.75	80%

step	12.51	8.48	17.24	**100%**
register(5-3)	-1.57	-3.58	0.26	**96%**
step:register(5-3)	5.39	-2.27	13.25	92%

Results of the All-cue Continuum show that the two cues, duration and glottalization, together clearly modulate the identification of the checked vs. unchecked contrast, we then examine the separate effects of duration and glottalization from results of Duration Continua, visualized in Figure 3.3. For the high register (Panel A), very little change in responses was observed as duration was manipulated in the two Duration Continua. In terms of the mid register (Panel B), however, increasing the duration (moving from left to right on the x-axis) resulted in a gradual increase in unchecked tone responses, similar to the pattern observed on the All-cue Continuum, in which all cues (both the spectrum and the duration) were proportionally manipulated. This seems to suggest that for the T3-T33 contrast but not for the T5-T55 contrast, duration plays an important role in the categorization of the checked versus unchecked tones.

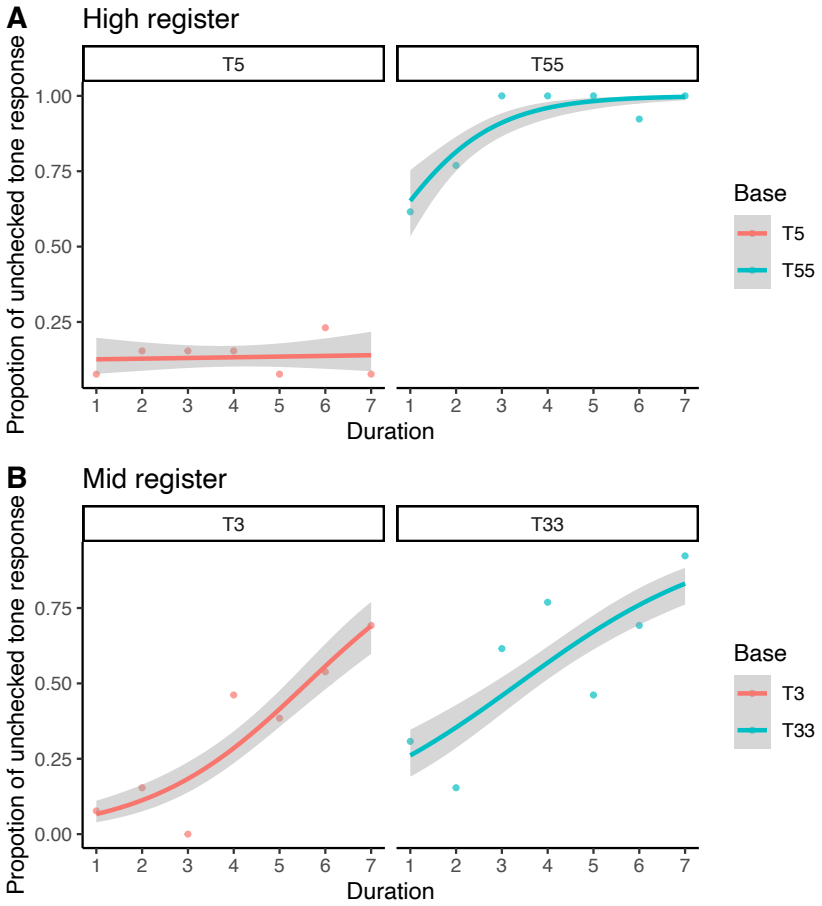


Figure 3.3: Empirical results of tone identification of the Duration Continua, grouped by tone register and colored by base.

To examine these observations from the visualization, we compare the relative contributions of glottalization and duration to the categorization with a Bayesian mixed effects logistic regression model on the data of the four duration continua. Predictors are duration, base, register, all the two-way interactions, and the three-way interaction. The predictor *duration* is centered and scaled using the *rescale()* function (Gelman, 2008) before model fitting. Predictors *base* and *register* are contrast coded (the checked base is coded as -0.5 and unchecked base is coded as 0.5; the mid register is coded as -0.5 and the high register is coded as 0.5). To account for individual-level variations, correlated by-participant random intercepts and random slopes of all the predictors (including two-way and three-way

interactions) are used. The regression model is fit to data by drawing 4,000 samples in each of four Markov chains from the posterior distribution over model parameters, with a warm-up period of 1,000 samples per chain. The following weakly informative priors are used: a normal distribution with a mean of 0 and standard deviation of 10 for fixed-effect terms, and LKJ prior with η=1.5 for correlations of random effects. Other parameters are fit with the default parameters.

Results of the model are presented in Table 3.2. First, all the simple terms, *duration*, *base* and *register* are credible predictors. Second, interaction terms other than the one between *duration* and *register* are credible. Specifically, it shows that the effect of *base* is credibly smaller for the mid register than for the high register (β = 9.79, 95% CI = [5.17, 14.86]), consistent with the visualizations in Figure 3.3. However, results in Table 3.2 can not provide details on the size of an effect at every given level of other predictors. For example, the effect size of duration for the continuum based on T5, which is one of our results of interest in Figure 3.3. We then inspect the effect sizes further using the *estimate_slopes()* function in the *modelbased* package (Makowski , 2020) and the *emmeans()* function in the *emmeans* package (Lenth, 2021). The *estimate_slopes()* function estimates the slope of a continuous variable's effect at given levels of another variable, here used to test the size (slope) of the duration at each level of *register* and *base*. The inspection reveals the following effect sizes of *base* for each register: (β = 2.66, 95% CI = [0.22, 5.50]) for the mid register and (β = 12.3, 95% CI = [8.81, 17.00]) for the high register. We inspect the effect of *duration* for levels of both *base* and *register*. Results reveal that duration is noncredible for the continuum based on T5 (β = 0.18, 95% CI = [-1.68, 2.24]), while it is credible for continua based on T3 (β = 6.51, 95% CI = [2.69, 10.86]), T33

(β = 5.16, 95% CI = [1.64, 9.45]) and T55 (β = 8.85, 95% CI = [3.61, 14.49]). Since these effects are rescaled by dividing by two times the standard deviations (Gelman, 2008), we can compare their effect sizes using the estimates. Consistent with the visualizations in Figure 3.3, for the high register, glottalization (i.e., *base*) is estimated to have a stronger effect than duration. Conversely, for the mid register, duration is estimated to have a stronger effect than glottalization.

Table 3.2: Results of the mixed-effect logistic Bayesian model for the categorization experiment. Probabilities of direction (*pd*s) of credible effects are bolded.

	β	l-95% CI	u-95% CI	pd
Intercept	0.86	-0.43	2.20	91%
duration	5.36	2.84	8.09	**100%**
base(unchecked-checked)	7.60	5.44	10.33	**100%**
register(5-3)	3.28	1.40	5.53	**100%**
duration:base(unchecked-checked)	3.47	-0.36	7.37	**96%**
duration:register(5-3)	-1.12	-4.47	1.80	76%
base(unchecked-checked):register(5-3)	9.79	5.17	14.86	**100%**
duration:base(unchecked-checked):register(5-3)	9.56	0.31	18.57	**98%**

3.2.4 Interim discussion

Through a 2AFC categorization paradigm, this experiment examined the role of vowel duration in perceiving the checked and unchecked tone contrast in TSM, for both the mid and the high registers. We found that for the mid tones, duration is a stronger cue than glottalization to identify checked vs unchecked tones, and that glottalization is a stronger cue than duration for the high tones.

These findings show that the role of duration for TSM checked-unchecked contrast perception is conditioned by tone register, the possible reasons of which will be discussed more in Section 3.4.1. Additionally, the fact that duration is indeed a reliable cue to the contrast in the mid register provides us the opportunity to examine the duration imitation mechanism. As reported above for the model presented in Table 3.1, estimates for the effect of duration were very similar for the T3-base and the T33-base continua, and furthermore, the credible intervals were highly overlapping. This suggests that the effect of duration on the perception of the contrast is quite similar for both continua. Considering that the presence of glottalization clearly indicates the checked tone, and thus may restrict duration imitation, and given that Pan (2017) reported that over 80% of final glottalizations were deleted in the production of checked tones, we chose to use the T33-based (unchecked) continuum in the imitation experiment to better match natural (unglottalized) productions of both tones.

In experiment 2, the same T33-based Duration Continuum will be used in an imitation task, to examine the linearity of duration imitation across the T3-T33 contrast.

3.3 Experiment 2: imitation

3.3.1 Method

3.3.1.1 Participant

To speed recruitment and ensure a high level of TSM proficiency, this experiment recruited participants from the TSM student club at several universities in Taiwan. Twenty participants (7 M, 12 F, mean age = 23, SD = 2) were recruited and 19 were analyzed after exclusion, with the criteria detailed in Section 3.3.3. The sample size follows that of Zhang et al. (2023)

and Kim and Clayards (2019), as the task is similar to those in the two studies; therefore a similar level of statistical power is needed.

3.3.1.2 Procedure

There were 3 tasks in this experiment. The first was a production task, the content of which were four TSM bi-syllabic words containing T3 or T33, repeated two times. These productions were used for evaluating the TSM proficiency of the participant by native listeners.

The second task was the 2AFC identification task, using identical instructions to Experiment 1. Considering that the participants may have a different (very likely higher) level of TSM proficiency from Experiment 1, we carried out the identification task as well to confirm the role of duration in the categorization. The stimuli, however, were half of that of Experiment 1 – only the mid register ones were tested in Experiment 2. Namely, the stimuli were four mid-register continua: Duration Continuum based on T3, Duration Continuum based on T33, 2 * All-cue continuum of the mid register. Each stimulus was repeated four times. Before the formal experiment, participants were given the same short instruction with audio examples and were provided the same 3 trials for practice as in Experiment 1.

The third task was the imitation task. Only the Duration Continuum based on T33 was used as the stimuli. Participants were instructed to repeat what they heard 'the closer to the way the speaker produced it the better'. There were five repetitions for each stimulus. Three trials were provided for practice.

The three tasks were blocked, and in each block the trials were randomized. Participants were able to take a rest after each block. As in Experiment 1, before the main task, the participants were asked to self-report their TSM proficiency through a questionnaire.

3.3.2 Statistical analysis

The identification results are analyzed using the same model and criteria as in Experiment 1 (Section 3.2.2). The imitation results are analyzed with the analogous analytical method as used in Zhang et al. (2024). As in that study: to determine if the imitation is more categorical or linear, two statistical models are fit and compared. The first statistical model is designed to reflect a production process where a talker categorizes the stimulus into one of the two phonological categories and then produces that category. Statistically this is modeled as a Gaussian mixture model with two Gaussian components (categories), and the probability of an imitation coming from one or the other is determined by the continuum step of the stimulus. The second model is designed to reflect a production process in which a talker tries to reproduce the phonetic properties as closely as possible with no role for categories. Statistically this is modeled as a linear model with only one Gaussian component but the mean and variance of that component vary according to the continuum step. We then compare the fit performance of the two models. As in Zhang et al. (2024), the models are compared using Pareto smoothed importance sampling leave-one-out cross-validation (PSIS-LOO CV), using the *loo* package (Vehtari et al., 2021). The first function, *loo()*, computes leave-one-out cross validation for each model. The model with a significantly higher expected log pointwise predictive density (ELPD) is considered a better fit for the data. The second function, *loo_model_weight()*, determines the weights of each model using the stacking method (Yao et al., 2018) in the optimal combination of the two models that best describes the data,

so-called *model averaging* (McElreath, 2015: Sec. 5.8). The model with the larger weight is interpreted as being able to better describe the data.

3.3.3 Results of the perception task

The production data was checked by a native TSM speaker (a collaborator on the current study), and all participants were judged as native-like. However, in the language questionnaire, one participant self-evaluated her TSM listening level as low as 1 in the 1-7 scale. To ensure the accuracy of the results, we excluded this participant's data.

The identification results of the All-cue Continuum are shown in Figure 3.4. Like in Figure 3.2, larger step indices increase the response rate of the unchecked tone. This increase is particularly pronounced in the middle range, with response rates approaching 0 and 100% for the first and last steps respectively, forming an 'S' shape identification curve. Figure 3.4 clearly demonstrates that participants in experiment 2 can distinguish between T3 and T33. As our focus is on the Duration Continua, we thus do not apply any statistical model to the results of the All-cue continuum.

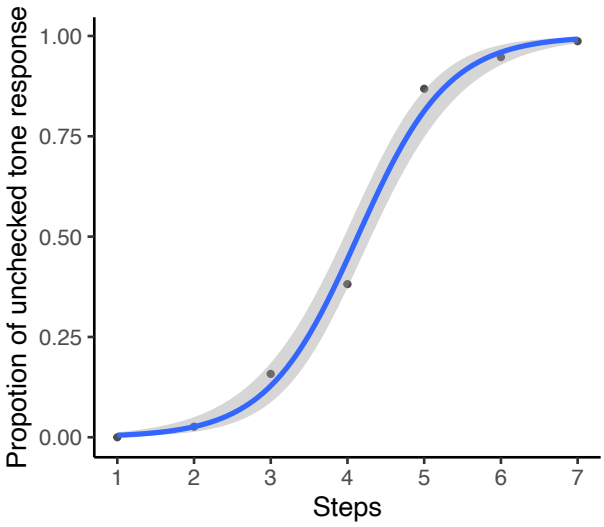


Figure 3.4: Tone identification results of the All-cue Continuum.

Figure 3.5 presents the identification of the Duration Continua. Analogous to Figure 3.4, longer duration increases the response rate of the unchecked tone and there is a pronounced increase. However, the response for the first and last steps is not always approaching 0 or 100%, indicating that the base token also plays a role in the categorization of the two tones. A Bayesian mixed-effect logistic regression model was fitted to the data of the Duration Continua. As in Experiment 1, the dependent variable is the probability of the unchecked tone response. *Duration* (centered and scaled), *base* (contrast coded: T3 is coded as -0.5 and T33 is coded as 0.5) and their interaction are used as the predictors. Correlated by-participant random effects are specified as random intercept and random slopes of all three predictors. The same priors, and sample sizes are used as the model presented in Table 3.1.

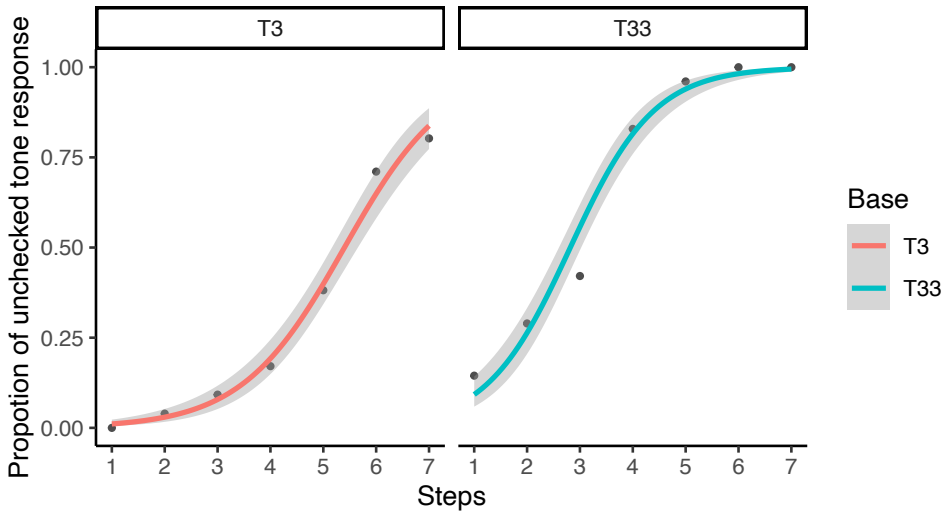


Figure 3.5: Tone identification results of the Duration Continua, grouped and colored by the base token.

Results of the Bayesian model are presented in Table 3.3. As expected, longer duration credibly increases the unchecked tone (i.e. T33) response ($\beta = 5.85$, 95% CI = [4.69, 7.15], pd = 100%). The continuum base of T33 also increases the T33 response compared to the base

of T3 (β = 3.47, 95% CI = [2.73, 4.30], pd = 100%), suggesting that the glottalization also plays a role in the categorization of T3 vs T33. Predictors are standardized before model fitting, allowing for comparisons of the estimates within the model. The estimate of *duration* is credibly larger than that of *base* (as their CIs do not overlap), indicating that duration is relied on more than glottalization by this group of participants in their categorization. This result replicates the critical role of duration for the mid-register tones in Experiment 1. Furthermore, the duration has a greater impact on the categorization of the T33-based Duration Continuum compared to the T3-based continuum (β = 1.63, 95% CI = [-0.04, 3.48], pd = 98%). This validates our decision to use the T33-based Duration Continuum in the imitation task.

Table 3.3: Results of the mixed-effect logistic Bayesian model for the categorization data. Abbreviations are the same as in Table 3.1.

	β	l-95% CI	u-95% CI	pd
Intercept	0.03	-0.45	0.49	53 %
duration	5.85	4.69	7.15	**100%**
base(T33-T3)	3.47	2.73	4.30	**100%**
duration:base(T33-T3)	1.63	-0.04	3.48	**97**%

3.3.4 Results of the imitation task

In processing the imitation data, target syllable boundaries were first automatically aligned using the Montreal Forced Aligner (McAuliffe et al., 2017, version 3.0.6), and then manually checked and corrected by the first author. The syllable onset was marked as the release of the

initial stop /l/ or the start of voicing, whichever occurred earlier. The offset was marked at the time that energy drops sharply. Syllable duration was measured between these two points.

Figure 3.5 shows density plots of the duration imitations as a function of the target duration values in the stimuli. In Figure 3.5A, the duration difference between the targets is well reflected in the imitations. Other than the two longest steps, there do not seem to be any merged distributions. Furthermore, in Figure 3.5B, the imitation variation (i.e., the width of the density distribution) increases with the increasing target duration, but there seems to be no increase in variation at the perceptual boundary step. In other words, the variation does not provide evidence of strong phonological mediation by the checked vs unchecked tonal contrast.

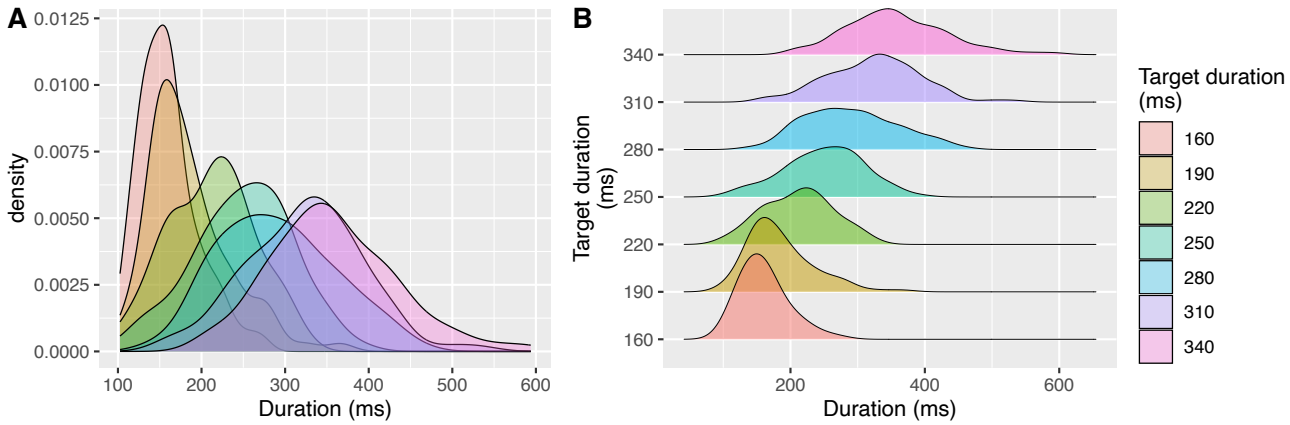


Figure 3.6: Density distributions of the imitated durations grouped and colored the by the target duration. Panel A: the superimposed density distributions. Panel B: the staggered density distributions.

As in Zhang et al. (2024), a uni-modal Gaussian model and a bi-modal Gaussian mixture model are fitted to the data, and then compared to determine which is a better fit. If the bi-modal model outperforms the uni-modal model, the data can be understood as

exhibiting a more categorical pattern. Conversely, if the uni-modal model outperforms the bi-modal model, the data can be considered more linear. The uni-modal model is structured by the following formulas,

duration ~ 1 + target_duration + (1+ target_duration | participant) (1)

sigma ~ target_duration + (1 + target_duration | participant)

(log link function) (2)

The imitated duration is the dependent variable and is used to fit the parameter *mu* (mean) of the Gaussian model in (1). The fixed effect *target_duration* allows the mean to vary with the target duration. By-participant random intercept and slope are also added to account for interspeaker variability. Formula (2) fits the *sigma* (log transformed by default), i.e., the standard deviation, of the Gaussian distribution. Similar to (1), the fixed effect *target_duration* allows the *sigma* to vary according to the target duration, as the visualization in Figure 3.5 shows that the width of the Gaussian distribution (*sigma*) increases with target duration. By-participant random intercept and slope are again fit. The predictor *target_duration* is centered and scaled before model fitting. Prior of the mean of the Gaussian is set to normal(mu=0.3 sec, sd=0.15), according to the visualization in Figure 3.5. The distribution of normal(0, 1.5) is used as the prior of fixed effect coefficients. For other model parameters, the default flat distributions are used: a half Student's t-distribution with 3 degrees of freedom and a scale parameter of 2.5 for random effect standard deviations, and a LKJ prior with $\eta = 1$ for correlations.

The mixture model is structure by the following formulas,

duration ~ 1, (3)

mu1 ~ 1 + (1 | participant), (4)

mu2 ~ 1+ (1 | participant), (5)

theta1 ~ 1 + target_duration + (1 | participant), (logit link function) (6)

sigma1 ~ 1 + (1 | participant), (log link function) (7)

sigma2 ~ 1 + (1 | participant), (log link function) (8)

In formula (3), the imitated duration is indicated as the dependent variable. Formula (4-5) fit the means of the two Gaussian mixtures, denoted in brms as *mu1* and *mu2*. Formula (6) fits the mixture probability *theta1*, which represents the probability that the imitated sample belongs to the first Gaussian mixture (i.e., the shorter tone T3) instead of the second (T33). In other words, *theta1* is the weight of the T3 category in the mixture; thus, (1-*theta1*) is the weight of the T33 category. Increasing the target duration in the stimulus is expected to decrease the probability of the imitated sample belonging to the first Gaussian component. Therefore, *target_duration* is used as a fixed effect for *theta1*, with by-participant random intercepts to allow for variability among participants (6). Because *theta1* is a probability, a logit link is used in the regression (6) by default so that only *theta1* values between 0 and 1 can be predicted. Formulas (7-8) fit the standard deviations of the two Gaussian mixtures, i.e., *sigma1* and *sigma2*. They are allowed to vary by participant, with no fixed effect. Like in (2), *sigma*s are log transformed in the model fitting by default.

Unlike previous models, some fixed effects have informative priors, to enable the model to converge on a unique solution (called "identifiability"), as is often necessary for non-linear Bayesian models (Bürkner, 2018). To ensure the first component has a shorter duration than the second component, priors are set as normal(0.15sec, 0.1) for *mu1* and normal(0.35sec, 0.2)

for *mu2* (based on the visualization in Figure 3.5). Following Zhang et al. (2024), a prior of normal(-15, 3) is used for the effect of target duration. For other parameters, weakly informative or default priors are used: cauchy(0.1,0.05) for *sigma*s, a flat distribution for the fixed-effect coefficient, a half Student's t-distribution with 3 degrees of freedom and a scale parameter of 2.5 for random effect standard deviations. To facilitate convergence, parameters were initialized at zero for every chain.

Both Gaussian models are fit to draw 4,000 samples in each of four Markov chains, with a warm-up period of 1,000 samples per chain. R-hat values are inspected to confirm convergence for both models. The results of the uni-modal model and the bi-modal model are shown in Table 3.4 and Table 3.5, respectively. For the uni-modal model, as expected, the mean of the Gaussian distribution increases credibly with the target duration (β = 0.14, 95% CI = [0.12, 0.15], pd =100%), and so does the sigma (β = 0.38, 95% CI = [0.24, 0.52], pd =100%). For the mixture model, as predicted, target duration shows a credible negative effect on theta1 (β = -17.46, 95% CI= [-21.98, -13.4], pd =100%) indicating that as target duration increases, imitations are less likely to belong to the first Gaussian distribution.

We then compare whether the uni-modal or the bi-modal model is a better fit to the data. Results of *loo()* show that the uni-modal model significantly outperforms the bi-modal model (ELPD = -163.2, SE = 19.7), suggesting that the imitation distribution is more linear than categorical. Nevertheless, this result does not mean that there is no categorical trace in the imitation at all. Following Zhang et al. (2024), we use model averaging to estimate to what extent the linear and categorical models jointly explain the data, using *loo_model_weight ()*. Results show that in the combined model that best describes the data, the weight for the uni-

modal model is 92% while that for the bi-modal model is 8%. The results further support the finding that the uni-modal model describes significantly more of the data.

Table 3.4: Summary of the unimodal Gaussian model for duration imitation

	Estimate	Est.Error	l-95% CI	u-95% CI	pd
Mean (Intercept	0.25	0.01	0.23	0.28	**100%**
sigma (Intercept)	-3.33	0.05	-3.44	-3.23	**100%**
Mean: target duration	0.14	0.01	0.12	0.15	**100%**
sigma: target duration	0.38	0.07	0.24	0.52	**100%**

Note: The sigma is log transformed in the model training. Same in Table 3.5.

Table 3.5: Summary of the Gaussian mixture model for F0 range imitation

	Estimate	Est.Error	l-95% CI	u-95% CI	pd
Mean of mixture 1 (Intercept)	0.18	0.01	0.17	0.20	**100%**
Mean of mixture 2 (Intercept)	0.32	0.01	0.29	0.34	**100%**
theta1 (Intercept)	-0.53	0.73	-2.04	0.86	77%
theta1: target duration	-17.46	2.19	-21.98	-13.40	**100%**
sigma1 (Intercept)	-3.39	0.11	-3.64	-3.20	**100%**
sigma2 (Intercept)	-2.97	0.07	-3.12	-2.83	**100%**

3.3.5 Interim discussion

When the value of a critical cue varies between two phonological categories, the imitation of the values of the cue tends to be non-linear, as has been observed in imitation of formants (Kim & Clayards, 2019), VOT (Flege & Eefting, 1988) and F0 (Zhang et al., 2023). However,

no investigations have been done so far for duration in cases where it plays a critical role to the imitated phonological contrast. Experiment 1 confirms the critical role of duration to the T3-T33 tonal contrast (but not for the T5-T55 tonal contrast) in TSM through a categorization experiment. Experiment 2 investigates both the categorization and the imitation of the T3-T33 contrasts by a new group of participants. The categorization results of Experiment 2 confirm the primary role of duration for the T3-T33 contrast. The imitation results of Experiment 2 reveal that the imitation of duration in the T3-T33 continuum is still linear, indicating that duration imitation appears not to be subject to phonological effects, in contrast to dimensions such as VOT, formant and F0. This finding suggests that duration possesses certain properties that allow it to bypass the phonological effect in imitation. Furthermore, it adds to our understanding that imitation is not universally driven by phonological mediation, but also by dimension-specific characteristics.

3.4 General discussion

This study explored the imitation mechanism of duration through two experiments. The first experiment investigated the role of duration in distinguishing checked vs. unchecked tonal contrasts in TSM, in relation to the role of glottalization. Three types of checked-unchecked continua were tested and compared in a 2AFC task, for both the mid and high registers. Results showed that duration was the most important cue for the T3-T33 (mid register) contrast but played a minor role for the T5-T55 (high register) contrast. In the second experiment, the T3-T33 contrast was used to examine the phonological mediation effect on duration imitation. Results demonstrated that the duration differences of the continuum steps

were well reproduced, and the imitation was much more linear than categorical. This finding suggests that the imitation of duration was not mediated by tonal categories, unlike other dimensions such as F0 and VOT.

3.4.1 The role of duration in TSM tones and implications for the onging tone merging

TSM tones received considerable attention in recent years on its sandhi rules and the merging of the two checked tones. However, less attention has been paid to how the checked vs unchecked contrasts are perceived. Phonologically, checked tones occur in syllables with one of four unreleased codas: /p, t, k, ʔ/ (Norman, 1988). However, the unreleased final coda is disappearing in TSM checked tones, at least for those with glottal stop (Pan 2017), suggesting the possibility that the weight of another cue, duration, may be enhanced in the perception of checked-unchecked contrasts. The first experiment of this present study thus examined to what extent listeners rely on duration to distinguish between checked tones and unchecked tones.

We found that the importance of duration in the checked-unchecked contrast perception is dependent on the tone register – it is a reliable cue for the T3-T33 contrast but not very important for the T5-T55 contrast. We provide some possible sources of why duration plays different roles in the perception of the checked-unchecked contrast for the two registers. First, Pan and Lyu (2021) conducted a study on TSM and found that when the coda [ʔ] was deleted, the preceding vowel was lengthened for the high tone (T5), causing a disconnection between the high checked tone and shorter durations. For the mid checked tone, however, the duration remained short after the coda deletion, which strengthened the connection between the mid checked tone and shorter durations. Therefore, findings in Pan and Lyu (2021) predict that

duration should play a smaller role in categorizing high checked vs. unchecked tones compared to mid checked vs. unchecked tones. Second, when creating the stimuli, we used naturally produced base samples for the two registers. As a result, the degree of glottalization was not strictly controlled across the registers. As shown in Figure 3.2, there seems to be more glottalization for 'la_T5' than for 'la_T3', which might bring about a larger glottalization effect in the Duration Continuum based on T5 compared to T3. However, the stimulus difference cannot account for the differing results between Duration Continua based on T33 and T55, as both continua were not expected to contain glottalization. Third, is the psycho-acoustic effect. A sound with higher pitch is perceived as longer than a sound with lower pitch, when their physical durations are equal (Gussenhoven & Zhou, 2013; Lu & Lee-Kim, 2021). As such, steps in the continuum based on T55 may have been perceived as longer than steps in the continuum based on T33, which may increase the unchecked tone responses overall, leading to ceiling effects. As shown in the right panel in Figure 3.3A, the shortest steps of the T55 Duration Continuum were often classified as checked, and even shorter syllables may have revealed a stronger effect of duration. Further investigations are needed to explore the effect of shorter durations on the checked-unchecked contrast perception for high register tones.

Results in this study may relate to ongoing sound change in the TSM tonal system. First, previous studies have suggested that the two checked tones (T5 and T3) are merging towards T3 and the final codas are undergoing loss (Ang, 2013; Pan, 2017), suggesting a decreasing phonological awareness of the checked-unchecked quality. Second, there seems to be an increasing awareness of durational quantity. As confirmed in the present study, the role of duration in distinguishing the checked vs. unchecked tones is pronounced for the mid register.

3.4.2 Imitation mechanism of duration

Substantial evidence has been found for the phonological effect in phonetic imitation in the past decades (Chistovich et al., 1966; Flege & Eefting, 1988; Kent, 1973; Kim & Clayards, 2019; Kwon, 2019; Repp & Williams, 1985; Zhang et al., 2023; Zhang et al., 2024). Similar to the perceptual non-linearity observed in segment categorization, these studies tend to find that the imitation of the critical cue is non-linear between the contrast it cues. The phonological effect and the non-linear imitation have been examined for dimensions including vowel formants (Chistovich et al., 1966; Kent, 1973; Kim & Clayards, 2019; Repp & Williams, 1985), VOT (Flege & Eefting, 1988; Nielsen, 2011), and F0 (Zhang et al., 2023, 2024). However, duration has not been investigated for the phonological effect in imitation as a primary cue for any phonological contrasts. Nevertheless, as a non-primary cue for phonological contrasts, and as a cue for speech prosody, duration imitation has frequently been observed in imitation tasks (Delvaux & Soquet, 2007; Giles et al., 1991; Kim & Clayards, 2019; MacLeod & Di Lonardo Burr, 2022; Pardo, 2010; Street & Cappella, 1989; Wagner et al., 2021; Zetterholm, 2007; Zhang et al., 2023). Therefore, it is worth examining if the commonly observed linear imitation of duration still emerges when it serves as a critical cue. This examination can add to our knowledge of whether the phonological mediation effect is dimension-specific or dimension-general.

Through an imitation experiment on a T3-T33 tonal continuum in TSM, this study found that duration was still linearly imitated although it serves as the critical cue to the contrast being imitated. On the one hand, this finding indicates a specific property of duration to be particularly sensitive in imitation. On the other hand, it shows that the phonological mediation effect is dimension-specific rather than dimension-general.

There are several possible reasons why duration has the special property of bypassing the phonological mediation effect. Firstly, duration plays multiple roles in speech prosody, such as marking word stress, focus prominence and prosodic boundaries. Similarly, duration is also involved in expressing para-linguistic (such as emotions) and non-linguistic (such as age) information (Pichora-Fuller et al., 2006; Scherer, 2015; Yildirim et al., 2004). The rich variations of duration in speech may lead listeners to be very sensitive in perceiving and producing the duration changes. However, this reason would also apply to other dimensions, such as F0, which is another dimension widely used in marking linguistic, prosodic, para-linguistic and non-linguistic functions. Nevertheless, the special property of bypassing the phonological mediation effect does not show up in the imitation of F0. This may be due to the differing nature of the processing mechanisms between duration and F0 (Liu, 2021). One example of such different mechanisms is that people with congenital amusia often sing out of tune but can sing on time (Dalla Bella et al., 2007). In addition, studies found that for both speech and music, timing information is processed in a more left lateralized way whereas the pitch information is processed in a more right lateralized way (Van Lancker Sidtis et al., 2021). Such differences in processing between F0 and duration could result in their behaving differently in phonetic imitation. Specifically, the phonological contrast could affect them to varying degrees, despite both of them being highly variable dimensions in human speech.

Secondly, the ability to control global speech rate may enable speakers to better perceive and reproduce duration changes. People can adjust their global speech rate based on various factors, such as cognitive demand (Nip & Green, 2013), social context (Street Jr & Brady, 1982), language proficiency (Temple, 2000), and cultural background (Verhoeven et al., 2004). During such adjustments, speakers should maintain the phonological contrasts, requiring

them to accurately produce the same phoneme in a wide range of durations. The rich practice of controlling the global speech rate could therefore be used to accurately reproduce the shortened and lengthened TSM tones. For instance, for checked tones, speakers may realize it with a range of durations across speaking rates that overlap with those of the unchecked tones. As a result, the phonological contrast does not lead to a discontinuity in accurately imitating the ambiguous steps between the checked and unchecked tones, although the non-linearity shows up in the categorization task.

The special characteristic of duration is also evident in other speech imitation and perception experiments. On the one hand, in experiments on convergence, which involve assessing the similarity of two productions by human listeners, results showed that timing relationships and duration, both of which are temporal in nature, were heavily relied upon by the listeners during the assessment (Pardo et al., 2013). Hence, the significant contribution of duration to the perception of phonetic similarity, as well as the precise reproduction of the duration in phonetic imitation, suggested the special characteristic/status of duration compared to other acoustic cues in speech. On the other hand, an early study investigating the perception of the English /i/ vs /ɪ/ contrast by native German, Spanish and Mandarin speakers revealed people's high sensitivity to duration in perceiving non-native vowel contrasts (Bohn, 1995). Although the three groups of participants differed in their use of duration as a cue in their native vowel phonology – duration is a cue for some German vowel pairs but not for Spanish or Mandarin vowels – they consistently showed a high reliance on duration rather than spectral cues when perceiving the English /i/ vs. /ɪ/ contrast. The results seem to reflect the overall perceptual salience of duration in speech perception. Taken together, these previous imitation and perception results, along with our findings in this study,

indicated listeners high sensitivity to perceiving duration and their high level of skill in producing accurate durations in speech.

3.4.3 Limitations and future directions

In the present study, we tested the same tonal continua of the mid register for two groups of participants in Experiment 1 and 2, respectively. Although results of both categorization experiments showed that duration is a more important cue than glottalization in the Duration Continua for the mid register, there were some differences in the results between the two groups of participants. The relative role of duration compared to glottalization for the T3-T33 contrast appears to be larger in Experiment 2 than in Experiment 1 based on visualization of the steepness of the curves in Figure 3.4 and Figure 3.3A. It is also suggested by the relative effect sizes of duration and base across the two categorization models. Although the median estimate of duration is larger than that of base for the mid register in Experiment 1, their CIs overlap a bit (duration: $\beta = 2.17$, 95% CI = [0.75, 3.8]; base: $\beta = 3.65$, 95% CI = [1.89, 5.66]). However, in Experiment 2, their CIs do not overlap. The discrepancy could be due to the differences in TSM proficiency between the two groups of participants. Although we did not evaluate the TSM proficiency of the participants for the first experiment, we recruited native speakers of TSM broadly from the university. For the second experiment, we recruited them from several TSM clubs, who were expected to have higher TSM proficiency levels than those in Experiment 1. Another possible reason is the smaller sample size (13) in Experiment 1 compared to Experiment 2 (19), which could have resulted in more noise in the statistical analysis in Experiment 1.

Another limitation of this study is that we did not control the degree of glotallization strictly in the T3 base and T5 base when creating the tonal continua; instead, we used the original productions directly from the speaker. There seems to be more glottalization in 'la_T5' than in 'la_T3', which may result in a larger glottalization effect in the Duration Continuum based on T5 compared to T3. However, as discussed earlier, the stimulus difference cannot account for the differing relative contributions of the two cues between Duration Continua based on T33 and T55, as neither of them contained glottalization, yet the two cues were utilized differently in the categorizations.

To obtain a more comprehensive understanding of the role duration plays in the perception of the checked vs unchecked tonal contrast for both registers, future investigations should explore larger ranges of duration variations, and in checked tones with other final stops. Additionally, it will be beneficial to investigate the individual-level links among checkedness perception, checkedness production and checkedness imitation to better understand tone merging processes and their implications for phonetic imitation.

3.4.4 Conclusion

In this study, we exmained the imitation of duration in a case where it plays an important role in distinguishing a tonal contrast. In the first experiment, we confirmed that duration is a reliable cue for the categorization of the mid-register checked vs unchecked tonal contrasts (i.e., T3-T33) in TSM. However, we found that duration is less important than glottalization in discriminating the checked vs unchecked contrast for the high register (T5-T55). In the second experiment, we examined the phonological effect in duration imitation using the T3-T33 contrast. We found that the imitation of duration was better fit by a unimodal model,

suggesting it is more consistent with a linear pattern that tracks the stimulus than mediation by the phonological (T3-T33) contrast. This finding demonstrates the unique property of duration in phonetic imitation – unlike other dimensions such as F0 and VOT, duration is not mediated by phonological contrast, whether it plays an important role or not. Furthermore, this finding indicates that phonological mediation in phonetic imitation is dimension-specific rather than dimension-universal, adding to our understanding of the mechanism of phonetic imitation.

References

Ang, U. (2013). The distribution and regionalization of varieties in Taiwan. *Language and Linguistics*, *14*(2), 315.

Boersma, P. (2020). Praat: Doing phonetics by computer [Computer program]. Version 6.1.16. *Http://Www. Praat. Org/*.

Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 279–304.

Brunelle, M. (2009). Tone perception in northern and southern Vietnamese. *Journal of Phonetics*, *37*(1), 79–96.

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411.

Chai, Y. (2021). Perception of checked tones in Xiapu Min. *The Journal of the Acoustical Society of America*, *150*(4), A309–A310. https://doi.org/10.1121/10.0008390

Chai, Y. (2022). Phonetics and Phonology of Checked Phonation, Syllables, and Tones [Ph.D., University of California, San Diego]. In *ProQuest Dissertations and Theses*. https://www.proquest.com/docview/2714115642/abstract/26EACB3B47A6499FPQ/1

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

Chen, S.-C. (2010). Multilingualism in Taiwan. *International Journal of the Sociology of Language*, *2010*(205). https://doi.org/10.1515/ijsl.2010.040

Chien, Y.-F., & Jongman, A. (2019). Tonal Neutralization of Taiwanese Checked and Smooth Syllables: An Acoustic Study. *Language and Speech*, *62*(3), 452–474. https://doi.org/10.1177/0023830918785663

Chistovich, L., Fant, G., de Serpa-Leitao, A., & Tjernlund, P. (1966). Mimicking of synthetic vowels. *Quarterly Progress and Status Report, Speech Transmission Lab, Royal Institute of Technology, Stockholm*, *1*(2), 1–18.

Coupland, J., Coupland, N., & Giles, H. (1991). Accommodation theory. Communication, context and consequences. *Contexts of Accommodation*, 1–68.

Dalla Bella, S., Giguère, J.-F., & Peretz, I. (2007). Singing proficiency in the general population. *The Journal of the Acoustical Society of America*, *121*(2), 1182–1189.

Delvaux, V., & Soquet, A. (2007). The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation. *Phonetica*, *64*(2–3), 145–173. https://doi.org/10.1159/000107914

Ding, P. S. (2016). *Southern Min (Hokkien) as a Migrating Language*. Springer Singapore. https://doi.org/10.1007/978-981-287-594-5

Flege, J. E., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. *The Journal of the Acoustical Society of America*, *83*(2), 729–740. https://doi.org/10.1121/1.396115

Garellek, M., Keating, P., Esposito, C. M., & Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America*, *133*(2), 1078–1089.

Giles, H., Coupland, N., & Coupland, I. (1991). 1. Accommodation theory: Communication, context, and. *Contexts of Accommodation: Developments in Applied Sociolinguistics*, *1*.

Gruber, J. F. (2011). *An articulatory, acoustic, and auditory study of Burmese tone*. Georgetown University.

Gussenhoven, C., & Zhou, W. (2013). Revisiting pitch slope and height effects on perceived duration. *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 1365–1369.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, *27*(6), 349–353.

Kent, R. D. (1973). The imitation of synthetic vowels and some implications for speech memory. *Phonetica*, *28*(1), 1–25.

Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, *34*(6), 769–786. https://doi.org/10.1080/23273798.2019.1582787

Kovaříková, A. (2023). *Subconscious Imitation of Phonetic Features of Perceived Speech and its Influence on Phonological Contrast*. Univerzity Palackého.

Kuo, C.-H. (2013). *Perception and acoustic correlates of the Taiwanese tone sandhi group* [PhD Thesis]. UCLA.

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), 10. https://doi.org/10.5334/labphon.83

Lin, M.-H. (2022). *Pitch and duration reflexes in Taiwanese Southern Min tones across generations.* [M.A.]. National Yang Ming Chiao Tung University.

Liu, X. (2021). Individual differences in processing non-speech acoustic signals influence cue weighting strategies for L2 speech contrasts. *Journal of Psycholinguistic Research*, 14.

Lu, Y.-A., & Lee-Kim, S.-I. (2021). The effect of linguistic experience on perceived vowel duration: Evidence from Taiwan Mandarin speakers. *Journal of Phonetics*, *86*, 101049. https://doi.org/10.1016/j.wocn.2021.101049

MacLeod, B., & Di Lonardo Burr, S. M. (2022). Phonetic imitation of the acoustic realization of stress in Spanish: Production and perception. *Journal of Phonetics*, *92*, 101139. https://doi.org/10.1016/j.wocn.2022.101139

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech 2017*, *2017*, 498–502.

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. https://doi.org/10.1016/j.cognition.2008.08.002

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142. https://doi.org/10.1016/j.wocn.2010.12.007

Nip, I. S. B., & Green, J. R. (2013). Increases in Cognitive and Linguistic Processing Primarily Account for Increases in Speaking Rate With Age. *Child Development*, *84*(4), 1324–1337. https://doi.org/10.1111/cdev.12052

Norman, J. (1988). *Chinese*. Cambridge University Press.

Pan, H. (2005). Voice quality of falling tones in Taiwan Min. *Ninth European Conference on Speech Communication and Technology*.

Pan, H. (2017). Glottalization of Taiwan Min checked tones. *Journal of the International Phonetic Association*, *47*(1), 37–63. https://doi.org/10.1017/S0025100316000281

Pan, H., Chen, M.-H., & Lyu, S.-R. (2011). Electroglottograph and acoustic cues for phonation contrasts in Taiwan Min falling tones. *Twelfth Annual Conference of the International Speech Communication Association*.

Pan, H., Huang, H., & Lyu, S. (2016). Coda Stop and Taiwan Min Checked Tone Sound Changes. *INTERSPEECH*, 1011–1015.

Pardo, J. S. (2010). *Expressing Oneself in Conversational Interaction*. 14.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, *69*(3), 183–195. https://doi.org/10.1016/j.jml.2013.06.002

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637–659.

Penney, J., Cox, F., & Szakay, A. (2018). Weighting of Coda Voicing Cues: Glottalisation and Vowel Duration. *INTERSPEECH*, 1422–1426.

Pichora-Fuller, M. K., Schneider, B. A., Benson, N. J., Hamstra, S. J., & Storzer, E. (2006). Effect of age on detection of gaps in speech and nonspeech markers varying in duration and spectral symmetry. *The Journal of the Acoustical Society of America*, *119*(2), 1143–1155.

Podlipskỳ, V. J., & Simácková, S. (2015). Phonetic imitation is not conditioned by preservation of phonological contrast but by perceptual salience. *ICPhS*.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria.

Repp, B. H., & Williams, D. R. (1985). Categorical trends in vowel imitation: Preliminary observations from a replication experiment. *Speech Communication*, *4*(1–3), 105–120.

Scherer, K. R. (2015). When and why are emotions disturbed? Suggestions based on theory and data from emotion research. *Emotion Review*, *7*(3), 238–249.

Street Jr, R. L., & Brady, R. M. (1982). Speech rate acceptance ranges as a function of evaluative domain, listener speech rate, and communication context. *Communications Monographs*, *49*(4), 290–308.

Street, R. L., & Cappella, J. N. (1989). Social and linguistic factors influencing adaptation in children's speech. *Journal of Psycholinguistic Research*, *18*(5), 497–519. https://doi.org/10.1007/BF01067313

Temple, L. (2000). Second language learner speech production. *Studia Linguistica*, *54*(2), 288–297.

Van Lancker Sidtis, D., Kim, Y., Ahn, J. S., & Sidtis, J. (2021). Do singing and talking arise from the same or different neurological systems? Dissociations of pitch, timing, and rhythm in two dysprosodic singers. *Psychomusicology: Music, Mind, and Brain*, *31*(1), 18–34. https://doi.org/10.1037/pmu0000270

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Vehtari, A., Gelman, A., Gabry, J., & Yao, Y. (2021). Package 'loo.' *Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. https://cran.r-hub.io/web/packages/loo/loo.pdf

Verhoeven, J., De Pauw, G., & Kloots, H. (2004). Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, *47*(3), 297–308.

Wagner, M. A., Broersma, M., McQueen, J. M., Dhaene, S., & Lemhöfer, K. (2021). Phonetic convergence to non-native speech: Acoustic and perceptual evidence. *Journal of Phonetics*, *88*, 101076. https://doi.org/10.1016/j.wocn.2021.101076

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Lee, S., Narayanan, S., & Busso, C. (2004). An acoustic study of emotions expressed in speech. *Eighth International Conference on Spoken Language Processing*.

Zetterholm, E. (2007). Detection of Speaker Characteristics Using Voice Imitation. In *Speaker Classification II* (Vol. 4441, pp. 243–257). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74122-0_19

Zhang, W., Clayards, M., & Sonderegger, M. (2024). *Qualitive differences in Mandarin tone imitation between Mandarin and English speakers*. LabPhon 19, Seoul.

Zhang, W., Clayards, M., & Torreira, F. (2023). Phonological mediation effects in imitation of the Mandarin flat-falling tonal continua. *Journal of Phonetics*, *101*, 101277. https://doi.org/10.1016/j.wocn.2023.101277

Preface to Chapter 4

The last two chapters investigate the phonological mediation effect on the imitation of F0 and duration, through a tonal contrast in Mandarin and another in Taiwanese Southern Min. Results showed that the phonological mediation varies depending on the type of dimension. Chapter 3 reveals that the imitation of duration is not mediated by tonal contrast, whereas Chapter 2 shows that of F0 imitation is indeed mediated by tonal contrast. In further exploring the phonological effect in phonetic imitation, Chapter 4 investigated the F0 imitation between speakers with different language backgrounds.

Flege and Eefting (1988) investigated how VOT was imitated by native English speakers, native Spanish speakers and English-Spanish bilinguals. They used a VOT continuum between 'da' and 'ta' as the stimuli and found that the three groups imitated the linearly changed VOTs into two or three categories. The categories differed across the three groups,

suggesting that the imitation of VOT was influenced by the phonological and phonetic categories associated with VOT in the participants' native language. However, few studies have investigated the imitation of phonetic continua by naïve speakers, i.e., speakers who lack experience with the phonological contrast upon which the phonetic continuum is based. Specifically, it is unclear if naïve speakers' imitation of F0 contours is linear or not, considering that they do not have lexical tone contrast in their phonology and thus will not be mediated by lexical-level phonological distinctions. Chapter 4 explores this question by examining the imitation of Mandarin tone continuum by naïve English speakers and comparing their results to those of native Mandarin speakers. The following predictions will be examined,

- If imitation of F0 is only mediated by lexical-level phonological contrasts, naïve English speakers will imitate the Mandarin tone continuum linearly.
- If imitation of F0 is mediated by categories from other levels (such as psychophysical or intonational ones), naïve English speakers will show categories in their imitation of the Mandarin tone continuum.

Chapter 4

Study 3: Imitation of F0 tone contours by Mandarin and English speakers is both categorical and continuous

4.1 Introduction

Imitation of manipulated stimuli has been found to be non-linear in many previous studies. When presented with a set of speech sounds in which one acoustic dimension is varied linearly (i.e., the difference of the dimension between two consecutive sounds is equal), imitators may reproduce the varying dimension in a non-linear manner (i.e., the differences in the imitated dimension between different pairs of consecutive sounds is unequal). For example, Nielsen (2011) investigated the imitation of /p/ with normal voice onset time (VOT) (72 ms), shortened VOT (30 ms), and extended VOT (113 ms), and found that the difference between normal and extended VOT was reflected in the imitation but the difference between normal and shortened VOT was lost. Several factors have been proposed to account for nonlinearity in phonetic imitation, including mediation by phonological contrasts. When a set of stimuli varies between phonologically contrastive sounds in the cue distinguishing these sounds, imitation tends to be more precise for values which are closer to prototypical values. For example, for the vowel contrast /ɛ/-/æ/ varying in formant frequency, imitation tends to be more precise for formant values which are closer to a prototypical /ɛ/ or /æ/, than for formant values that are intermediate or ambiguous between the two vowels (Kim & Clayards,

2019). Likewise, in the imitation of sounds between /b/ and /p/ along a VOT continuum, linearly varied VOTs were reproduced in clusters, with the centers aligning with the prototypical /b/ or /p/ in the imitators' native phonology (Flege & Eefting, 1988). Such nonlinearity in imitation related to phonological contrasts has also been found in the imitation of F0 contours between Mandarin flat and falling tones (Zhang et al., 2023), and between early peaked and late peaked rise-fall-rise intonations in English (Pierrehumbert & Steele, 1989).

While this phonological mediation has been observed in phonetic imitation, less is known about potential mediation by non-native contrasts. In addition, although speech perception has been widely studied in the past decades, the precise relationship between phonetic imitation and speech perception—specifically how imitation aligns with or is restricted by perception—remains not fully understood. This study aims to shed light on these issues by comparing the imitation and perception of Mandarin lexical tones, between native Mandarin speakers and English speakers who are naïve to Mandarin.

Furthermore, this study uses a statistical method that builds on our previous work for modeling imitation in terms of phonological categories (Zhang et al., 2023): a "mixture regression model", in which speakers are modeled as transitioning between categories as a phonetic cue is varied. The mixture model is compared to a model where listeners are simply tracking the acoustic cue (without underlying categories), which is implicitly assumed in most prior work. We find that the data is best fit as a weighted average of these two models, suggesting that both are necessary in accounting for imitation variations. This leads to the conclusion that both pre-exiting categories and within-category cue tracking are involved in phonetic imitation.

4.1.1 Links between perception and imitation

The process of imitation is generally thought to involve three phases: perception, storage in memory, and reproduction (Flege & Eefting, 1988). Phonological mediation may take place during the perception phase, as phonological contrasts could warp listeners' perceptual space (Iverson & Kuhl, 1995; Liberman et al., 1957; but see also McMurray et al., 2008). As a result, any non-linearity in the perception phase could be carried over to the reproduction phase. In speech perception tasks requiring categorization, listeners also show non-linearities, tending to respond to differences between categories and sometimes showing reduced sensitivity within a single category (Liberman et al., 1957). This pattern is seen in two types of task performed on the speech continuum between two categories: identification tasks and discrimination tasks. In an identification task for speech categorization, the identification curve of one category in the contrast will be sigmoidal, with a sharp increase in the response rate indicating the perceptual boundary of the two categories. In the discrimination task, the discrimination accuracy tends to peak at the category boundary position, but remains closer to chance level within each category. It is important to note that these results don't necessarily mean that perception is only categorical (and this pattern is not always observed, see Pisoni, 1973), but rather that behaviour can show effects of categories (Pisoni & Tash, 1974; see McMurray, 2022 for a review).

The values of the dimension that are poorly imitated seem to correlate with the category boundary position in speech categorization. Kim and Clayards (2019) investigated the perception and imitation of English /ɛ/-/æ/ continua by native speakers. They found that for the perceptually ambiguous steps (in terms of the /ɛ/ or /æ/ category), the imitation was less accurate, with higher variance in F1-F2 space. Similarly, Zhang et al. (2023) carried out

perception and imitation experiments on Mandarin flat-falling tonal continua for native speakers. As in Kim and Clayards (2019), they found that the imitation of the intermediate two F0 steps (out of a total of seven F0 steps) exhibited inconsistency and high variance. Furthermore, they observed two distinct modes in the distribution of the F0 range imitation, which corresponded to the two tonal categories. The valley between the two modes aligned with the category boundary position of the tonal contrast, as observed in the perception results. A similar boundary effect has also been observed in the imitation of intonational contours. Pierrehumbert & Steele (1989) investigated the imitation of rise-fall-rise contours in English, where the timing of the F0 peak was manipulated from early to late. They observed two distinct modes in the distribution of the imitated peak timing, and a valley between the modes corresponding to the categorical boundary between two English pitch accents that contrast in peak timing: L*+H (early peak) and L+H* (delayed peak). These findings suggest that categorization effects are evident in both imitation and perception for segmental and intonational contrasts, with the category boundary in imitation experiments aligning with the boundary observed in perceptual categorization experiments.

4.1.2 Psychophysical boundary vs. linguistic boundary in categorization

Studies have found that language experience affects speech categorization, raising the question of how language experience would impact imitation. Perception studies on tonal continua between Mandarin tonal contrasts have found that native speakers' perception of Mandarin tones shows category effects (Wang, 1976), in both identification and discrimination experiments. However, recent studies have found that within category details were not entirely lost (Gao et al., 2019; Qin et al., 2019), in line with findings on other tonal

contrasts. For non-native speakers whose native language is tonal (e.g., Cantonese), the perception pattern of Mandarin tones was also evident in both identification and discrimination experiments, although they differed from native Mandarin speakers in the position and steepness of the category boundary (Peng et al., 2010). Interestingly, for non-native speakers whose native language is non-tonal (English and German), the categorization pattern was observed in identification experiments but not in discrimination experiments (Peng et al., 2010; Shen & Froud, 2016; Wang, 1976; Xu et al., 2006). Wang (1976) proposed that two boundaries co-exist in speech categorization: the psychophysical boundary and the linguistic boundary, corresponding to the boundaries in the identification experiment and in the discrimination experiment. The psychophysical boundary indicates that listeners, in general, can distinguish rising or falling from flat F0 contours. In contrast, the linguistic boundary indicates that listeners' perceptual sensitivity to F0 contours is shaped by the tonal categories in their native phonology. In addition, a linguistic boundary can be developed through Mandarin learning experience. Shen and Froud (2016) found that native English speakers who learned Mandarin as an L2 showed both boundaries in perception experiments on Mandarin tones, while those without prior Mandarin experience only showed the psychophysical boundary.

In a study of Mandarin tone imitation by native speakers, Zhang et al. (2023) observed tonal boundaries in the distribution of imitated F0 ranges. However, it remains unclear whether the observed boundary in native imitation was the psychophysical boundary, the linguistic boundary, or both. In other words, it is unclear whether phonetic imitation is mediated by linguistic categories, psychophysical categories, or a combination of the two. This study investigates this question by examining the effect of language experience on

Mandarin tone imitation by two groups of participants: native Mandarin speakers and naïve speakers whose native language is non-tonal. If imitation is mediated by psychophysical categories, which are independent of participants' native language, we expect to observe categories in the imitation of both native Mandarin speakers and naïve non-tonal speakers. On the other hand, if imitation is mediated by linguistic categories, we expect that categories would not be present in the imitation of naïve non-tonal speakers.

To address this question, it is necessary to investigate tone imitation by naïve speakers of Mandarin who have no tonal background. As such, this study recruited native English speakers with no previous tonal experience (henceforth, naïve English speakers) as participant. This approach allows us to isolate the effect of native language on Mandarin tone imitation, without the confounding influence of pre-existing linguistic boundaries.

It is worth noting, however, that studies have found that intonational categories can mediate phonetic imitation (Pierrehumbert & Steele, 1989; Cole et al., 2023), raising the possibility that English speakers may employ English intonational categories when imitating Mandarin tones. As detailed in Section 4.1.4, we adopt the flat versus falling tonal contrast in Mandarin for this study, which raises the question of whether a potential flat-falling intonational contrast is involved in English speakers' imitation of the Mandarin flat-falling tonal contrast. However, we follow previous studies in assuming that the perception of Mandarin flat-falling contrast for English listeners is primarily non-phonological (Chang et al., 2016; Hallé et al., 2004; Peng et al., 2010; G. Shen & Froud, 2019; Xu et al., 2006). We outline our reasons bellow.

First, second language assimilation experiments have shown that the flat tone and the falling tone in Mandarin were not consistently mapped to different English intonational

categories. So and Best (2008) proposed four intonation categories in English that could correspond to the four lexical tones of Mandarin: *Flat pitch* (targeting the flat tone), *Question* (rising), *Uncertainty* (dipping), and *Statement* (falling). They tested these correspondences by examining how naïve Australian English speakers assimilated (categorized) the Mandarin tones (in citation form) into these four intonational categories. They found that the falling tone was assimilated primarily to the *Statement* category while the flat tone was assimilated to both the *Statement* and *Flat pitch* categories. When So and Best (2011, 2014) embedded the Mandarin tones in a carrier sentence, they found that both the flat tone and the falling tone were assimilated primarily to the *Statement* category, indicating that the two Mandarin tones were not consistently mapped to different intonational categories.

Second, it is unclear if *Flat pitch* is a well-motivated intonational category. Steffman et al. (2024) used imitation to examine the underlying categories for 12 surface F0 forms in English, including flat and falling tunes[3]. They found that the flat tunes did not form a separate category but were instead merged into the rising or falling category, and were not produced with high pitch. This suggests that a high flat F0 contour, like the flat tone in Mandarin, may not correspond to an intonational category for native English listeners.

Third, Shen and Froud (2019) found that English speakers did not show native-like automatic categorization of the Mandarin flat-falling contrast in their electrophysiological response. Instead, they showed high sensitivity to both cross-category and within-category F0 deviants. Taken together, these findings raise doubt that intonational contrasts in English

[3] Here the flat tunes refer to representations of H*HL and LH*HL in Figure 1 in Steffman et al (2024). Falling tunes refer to H*LL, LH*LL and L*HLL in the same figure.

directly mediate the perception of the Mandarin flat-falling contrast. We return to this question in the discussion.

4.1.3 Imitation studies of Mandarin tones

Previous studies have provided insights into the perception and production of Mandarin tones by naïve English speakers. Specifically, findings from perception studies have shown that naïve English speakers attended more to F0 height than to F0 shape when perceiving Mandarin tones, compared to native Mandarin speakers (Francis et al., 2008; Gandour, 1983; Huang, 2004). Furthermore, a perceptual training experiment on Mandarin tones found that successful English learners of Mandarin tones attended more to pitch direction than less successful English learners (Chandrasekaran et al., 2010). Taken together, these results indicate that naïve English speakers are less sensitive to F0 slope or range of Mandarin tones compared to native Mandarin speakers, and tend to prioritize F0 height.

Bent (2005) examined the perception and production of Mandarin tones in monolingual English speakers. Since the participants had no experience with lexical tone, the production experiment used an imitation paradigm, where participants imitated canonical Mandarin tones produced by a native male speaker. In general, participants performed better in the imitation task than in the perception task. The tone shapes imitated by the monolingual English speakers were distinguishable, and none of the tones were collapsed in the imitated productions. Nevertheless, their imitations differed from those of a comparison group of native Mandarin speakers. The imitated tones exhibited a compressed pitch range, and the starting and ending points, as well as peaks and valleys in the imitated contours all showed positional differences compared to the imitations by native Mandarin speakers. Two further

studies also adopted the imitation paradigm to investigate Mandarin tone acquisition for English learners (Hao, 2012; Hao & de Jong, 2016). Following Bent (2005), they used canonical Mandarin tones as the imitation material and used native Mandarin speakers' judgements to evaluate imitation accuracy. Like Bent (2005), Hao (2012) and Hao and de Jong (2016) found that participants performed better in imitation than in perception tasks. However, the latter two studies did not evaluate the contours of the imitated tones.

These studies suggest that imitation of Mandarin *canonical* tones by native English speakers is not very difficult, although they may show a reduced F0 range. However, to the best of our knowledge, no work has investigated monolingual English speakers' imitation of F0 contours both within and across Mandarin tonal categories. It remains unknown whether they can imitate the degree of F0 fall between the canonical flat and falling tones as accurately as they imitate the canonical tones themselves.

4.1.4 The current study

The main goal of the present study is to explore how the different types of boundaries, i.e., linguistic vs. psychophysical, affect the imitation of F0 contours. We compare the imitation of a Mandarin tonal contrast between native Mandarin speakers and naïve English speakers. To examine how imitation is related to perception, categorization experiments are carried out on the same continua with a different group of naïve English and native Mandarin speakers.

Following Zhang et al. (2023), we investigate imitation of stimuli along an F0 continuum between the Mandarin flat and falling tones. We analyze the imitation data through both Gaussian mixture regression models and simple Gaussian regression models. As will be detailed in Section 4.4.3, the mixture regression model assumes that speakers are

transitioning between categories as a phonetic cue is varied, whereas in the simple Gaussian model, speakers are assumed to simply track the F0 variation without the effect of underlying categories. Conclusions will be drawn from the performance of the two models in fitting the imitation distribution. First, if the simple Gaussian model greatly outperforms the mixture regression model, it suggests that speakers are tracking the F0 cue phonetically, and imitation is not affected by pre-existing categories. Second, if the mixture regression model greatly outperforms the simple Gaussian model, it indicates that pre-existing categories are mediating the imitation of F0s, and phonetic cue tracking is not an important factor. Additionally, results in Zhang et al. (2023) indicate a third possibility. Although Zhang et al. (2023) found that the mixture model was a better model than the simple Gaussian model, the difference in fit was not large. Native speakers did reproduce the differing degrees of F0 fall within the falling tone category, indicating the possibility that categorical mediation and phonetic tracking co-exist in imitation behavior. As detailed below, we will examine this possibility by considering a weighted combination of the two models to explain the imitation data.

4.2 Materials

Following Zhang et al. (2023), the syllable 'ba' was used to carry the tonal continuum. A new base syllable was produced in isolation by the same female Mandarin speaker as in the previous study. A 9-step tonal continuum was created in Praat following the same procedure as before. The F0 onset, F0 offset and the F0 falling range of each step are presented in Table 4.1. The F0 values of the two end points (step 1 and step 9) were determined based on the averaged values of five isolated productions of the flat tone and the falling tone by the speaker.

The duration of the continuum was 330 ms, which was the mean duration of these same productions.

Each stimulus consisted of the isolated syllable 'ba' carrying one of the F0 continuum steps. This is different from Zhang et al. (2023), where a carrier sentence ('ba de0 sheng1', meaning 'the sound of ba') was used as the imitation stimulus. We used isolated syllables to ensure that English-speaking participants would not face difficulties posed by the Mandarin carrier sentence. The stimuli, data and analysis code for this study are available on an open-access OSF repository at https://osf.io/6q3je/.

Table 4.1: Onset, Offset and F0 falling range values for each F0 range step

Step	1	2	3	4	5	6	7	8	9
Onset (Hz)	240	244	248	251	255	259	263	266	270
Offset (Hz)	240	234	228	221	215	209	203	196	190
F0 Range (Hz)	0	10	20	30	40	50	60	70	80
F0 Range (st)	0.0	0.7	1.5	2.2	3.0	3.7	4.5	5.3	6.1

4.3 Perception experiment

The identification experiment had two aims. First, it aimed to determine the perceptual boundary within the flat-falling tonal continuum within each group, to investigate whether the distribution of imitations is linked to perceptual categories. Second, it sought to compare the perceptual boundary between native Mandarin speakers and naïve English speakers.

4.3.1 Participants

We recruited 15 Mandarin participants for the categorization experiment from Prolific, an online recruitment platform. They all reported 'Chinese' as their first language and most fluent language to Prolific. However, on checking the language questionnaire included in the experiment (as introduced in the following section), we found that three participants reported speaking Cantonese as their native language[4]. They were excluded from the analysis, leaving 13 Mandarin speakers (mean age: 28.2, standard deviation: 5.3, 6M 6F) for analysis. Five of the Mandarin speakers reported growing up in the northern part of China, seven reported growing up in the southern part of China. We should note that although remote data collection offers convenience, it comes with decreased control over participant background. In particular, we did not control for regional varieties of Mandarin that the participants may speak in addition to Standard Mandarin (*Putonghua*), which is in fact not the first language of most Chinese speakers. We assume in this study that everyone educated in the education system in China can be seen as a 'native speaker' of standard Mandarin (henceforth native Mandarin speaker), as they begin learning Standard Mandarin from a very young age and default to reading and speaking it in most contexts, which we assume includes this experiment. When collecting the English data on Prolific, we selected participants who reported themselves to be monolingual speakers of English, born and currently residing in Canada or the US. We recruited 13 English speakers (mean age: 34, standard deviation: 9, 8M 5F) for this experiment. Their language questionnaire responses were checked to ensure that these

[4] This discrepancy was most likely due to the fact that, at the time the experiment was conducted, Prolific did not list 'Cantonese' as an option when collecting language information from the participants, whereas our language questionnaire did. As a result, the three participants had to report 'Chinese'-the closest language to Cantonese-as their first and most fluent language to Prolific, but select 'Cantonese' on the language questionnaire.

participants had no tonal experience. These experiments were set up using the prosodylab experimenter (Wagner, 2021), a set of javascripts building on jsPsych (De Leeuw, 2015).

4.3.2 Procedure

Participants first filled in the consent form and language questionnaire, then were directed to the main experiment. The two-alternative forced choice paradigm was employed in the identification experiment. In each trial, a stimulus representing one of the 9 steps in the continuum was played, and participants were presented with two options on the screen. For English participants, both PINYIN form and word description of the tones were provided in the options: "the flat tone (bā)" or "the falling tone (bà)" . For Mandarin participants, only the PINYIN form was presented. Participants were instructed to press one of two number keys to make their decision: '1' for the flat tone, '4' for the falling tone. They had unlimited time to respond.

For English participants, there was a training session and a practice session prior to the main experiment. The training session included two good examples of each tone, which were presented with the labels described above. In each example, participants could play the audio of the tone an unlimited number of times (they were not able to replay audio in any other section of the experiment). They then performed three trials in a practice session to further familiarize the participants with the task. There was no training session or practice session for Mandarin participants.

In the main experiment there were 5 repetitions for each step of the continuum, and the stimuli were presented in a random order. The experiment took approximately 7 minutes for Mandarin participants and 10 minutes for English participants.

4.3.3 Results

The identification curves of both groups of participants are shown in Figure 4.1. Mandarin listeners show a sigmoidal response pattern, with response rates close to 0 or 1 for extreme steps and a sharp change of response rate in the middle of the continuum (around 30 Hz). In contrast, the response rates for English speakers at extreme steps do not reach to 0 or 1, although the 50% crossover seems also to be around 30 Hz.

We fit a Bayesian mixed-effect logistic regression model to the response data (falling tone =1, flat tone = 0) to identify the category boundary position and assess the steepness of the curve. Fixed effects included *F0 range* (centered and divided by 2 SD), participant *Group* (dummy coded, EN: 0, MD: 1), and the interaction of the two. The random effects included by-participant intercepts and slopes for *F0 range*. The model was fit by drawing 4,000 samples in each of four Markov chains from the posterior distribution over model parameters, with a warm-up period of 1,000 samples per chain, retaining 75% of samples for inference. All Bayesian regression models in this study were implemented with *brms* (Bürkner, 2018; version 2.18.0), an R-based front-end to the Stan programming language (Stan Development Team, 2019, 2021). The model was fit with default priors: a flat distribution for fixed effect coefficients, a half Student's t-distribution with 3 degrees of freedom and a scale parameter of 2.5 for intercepts and random effect standard deviations, and a LKJ prior with $\eta = 1$ for correlations.

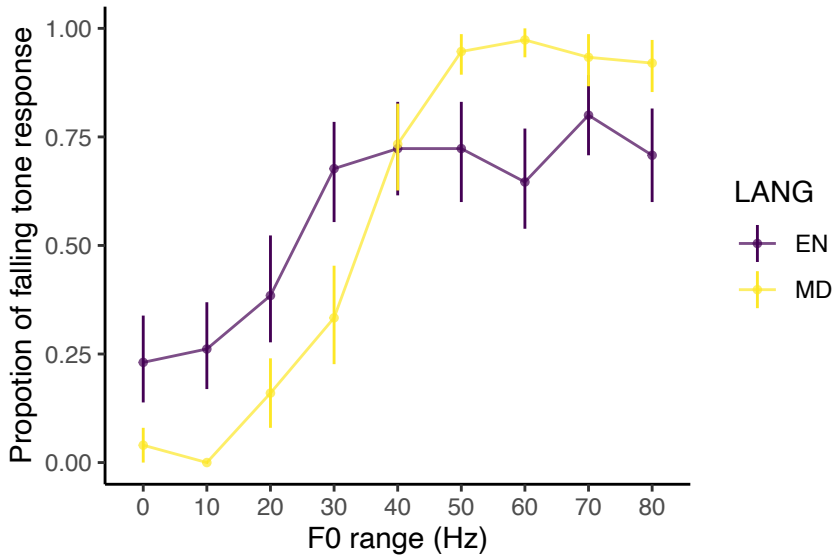


Figure 4.1: Rate of response of the falling tone as a function of the F0 falling range (EN for English speakers and MD for Mandarin speakers, the same in following figures).

The model's fixed effects are summarized in Table 4.2, where the median and standard error for an estimate's posterior as well as 95% credible intervals (CI) are reported. A CI excluding 0 (i.e. no estimated effect) can be taken to provide compelling evidence for an effect (Nicenboim & Vasishth, 2016). We also report the probability of direction (PD), computed with *bayestestR* (Makowski et al., 2019). A PD is the percentage of a parameter's posterior distribution with a given direction, and we interpret a PD higher than 95% to indicate that an effect is credibly present. The estimate for the *Group* by *F0 range* interaction showed that the English response curve was much shallower than that of Mandarin participants (estimate = 6.87, 95% CI = [3.57,10.82], PD = 100%). Thus, English speakers exhibited a less clear-cut categorical distinction in their perception of the falling tone compared to Mandarin speakers. To test if the boundary positions of the two groups are different from each other, we used the model to compute the posterior distributions of the crossover point for *Group* = EN and the crossover point for *Group* = MD. The estimated crossover points were 34.4 Hz (i.e., between

the third and fourth steps) for MD speakers (95% CI = [29.7, 39.0]) and 28.4 Hz (i.e., between the second and third steps) for EN speakers (95% CI = [-16.4, 51.4]).

Table 4.2: Results of the mixed-effect logistic Bayesian regression model of the response rate of the falling tone (fixed effects only, CI: Credible Interval, PD: Probability of Direction). A PD indicating a credible effect is bolded. Same in Table 4.4-4.5.

	Estimate	Std.Error	l-95% CI	u-95% CI	PD
(Intercept)	0.59	0.44	-0.26	1.50	92%
F0 range	2.55	1.17	0.30	4.93	**99%**
Group (MD-EN)	0.42	0.64	-0.82	1.73	75%
F0 range*Group (MD-EN)	6.87	1.84	3.57	10.82	**100%**

4.4 Imitation

4.4.1 Participants

We recruited 17 Mandarin participants who reported 'Chinese' as their first language and fluent language on Prolific. The sample size follows that of Zhang et al. (2023) and Kim and Clayards (2019), as the task is similar to those in the two studies, which implies that a similar level of statistical power is needed. Native speaker status was checked by a native Mandarin speaker (the first author) by listening to a recorded reading passage (which will be described in Sec 4.2). We consider all individuals who have acquired Standard Mandarin through the education system in China from a young age as native speakers of Mandarin. This is subject to the same caveat raised in 4.3.1., that they may also speak another regional variety of

Mandarin. Six Mandarin speakers were excluded as they had a clear non-native accent. Another Mandarin speaker was excluded because of severely creaky productions. As a result, there were 10 Mandarin speakers (2M 8F) in the final sample that were recruited from Prolific. To ensure a balanced gender distribution and an equal number of participants for both languages, as well as considering the relatively high rate of non-native Mandarin speakers on Prolific, we recruited another 7 male Mandarin speakers through a social media post. Thus, we analyzed data from 17 Mandarin participants (mean age: 26, standard deviation: 4, 9M 8F). Nine of the Mandarin speakers reported that they grew up in the northern part of China, seven reported growing up in the southern part of China, and one reported growing up in Malaysia. Nineteen English participants (mean age: 36, standard deviation: 10, 10M 9F) were recruited from prolific, and we required that speakers report that they were monolingual speakers of English, born and currently residing in Canada or the US. We also checked their language questionnaire to ensure that they have no experience with tone languages.

4.4.2 Procedures

Participants first filled in the consent form and language questionnaire, then were directed to the main experiment. There were two tasks in the experiment. The first was a reading task in which participants read the short passage 'The North Wind and The Sun'. The Mandarin version of the passage contained approximately 160 Chinese characters, while the English version was around 110 words. Participants were instructed to read the short passage aloud after a silent reading warm-up. The production data from this task was used to estimate the speaking F0 range of each participant. Since tonal targets are not absolute F0 values but relative pitch heights, we used normalized F0 within the speaker's F0 range when comparing the pitch trajectories across speakers. According to Baken and Orlikoff (2000), in English, a

speaker's speaking F0 range can be estimated from 3-4 sentences. We used a short passage to obtain a more precise estimate. The second task was the imitation task. Participants were instructed to imitate the stimuli as closely as possible to the way they were produced by the speaker. Each stimulus was repeated five times and all trials were presented in a randomized order.

4.4.3 Results

Target syllable boundaries were first automatically obtained using the Montreal Forced Aligner (McAuliffe et al., 2017, version 3.0.6), then manually checked and corrected by the first author. F0 was extracted using Praat (Boersma, 2021), with the following procedures to minimize F0 tracking errors. We first extracted the F0 of the reading passage for each speaker using the default settings, then analyzed the F0 distribution – the median F0 in Hz (50% quantile) as well as the 25% and 75% quantiles were obtained for each participant. Second, we calculated each participant's speaking F0 range using formulas from De Looze and Hirst (2008): ceiling = $F0_{(75\%\ quantile)} * 1.5$; floor = $F0_{(25\%\ quantile)} * 0.75$. Third, for each speaker, their individual ceiling and floor values were used as parameters in Praat to extract the F0 values of their imitations. In each syllable, 10 F0 measurements were extracted, each of which was the mean F0 within 10% of the syllable. We found this method worked well in reducing F0 tracking errors compared to Praat's default settings. To normalize the F0 ranges across speakers, we transformed Hz to semitones (st) using the formula in (1), in which the F0ref is the participant's mean F0. F0 range was calculated in the following steps: First, the maximum F0 in the first 50% of the target syllable was obtained, denoted as $F0_{Max\_1stHalf}$. According to visualization of the pitch tracks and experience from previous studies (Wang et al., 2020; Xu, 2005), the Max F0 of a falling tone in general happens in the first half of the syllable, whereas

the Min F0 typically occurs near the end of the syllable. Second, the minimum F0 was calculated as the last F0 extracted, i.e., the mean F0 in the last 10% of the syllable, denoted as $F0_{Min\_final}$. In cases where the F0 of last 10% was not extracted successfully, we used the preceding F0 measurement as the $F0_{Min\_final}$. Finally, the F0 range was calculated as $F0_{Max\_1stHalf} - F0_{Min\_final}$.

$$st = 12 * log_2(\frac{F0}{F0_{ref}}) \quad (1)$$

Smoothed F0 trajectories of the imitated F0 contours (after normalization) for both languages are shown in Figure 4.2. Time normalized measurements were converted to msec in order to visualize the differences in length of productions as well as their F0 contours. To facilitate easier comparisons of offset values among the steps, the F0 contours are aligned to the end of the syllable. Figure 4.2 suggests that English speakers accurately reproduced the overall falling contours, despite more variation at the beginning of the contours. Both English and Mandarin speakers show an increase in the imitated F0 range as the target F0 range increased, indicating their sensitivity to fine F0 variations. Both findings align with the results reported by Zhang et al. (2023). However, the imitated F0 contours appear 'more continuous' for English speakers than for Mandarin speakers. The Mandarin speakers (right panel) seem to show a separation between steps 1-4 and steps 6-9, while the English speakers (left panel) show less distinction between these steps. Additionally, English speakers imitate the F0 contours with a very different average pitch height from Mandarin speakers. Note that the st values are normalized according to each speaker's mean F0 in their passage reading. Consequently, positive st values indicate higher pitch than the speaker's mean F0. English speakers use a higher overall pitch level (relative to their mean F0) than native Mandarin speakers. Furthermore, Mandarin speakers seem to produce the falling tones with a shorter

duration compared to flat tones, with steps 6-9 shorter than steps 1-4. English speakers do not show this pattern.

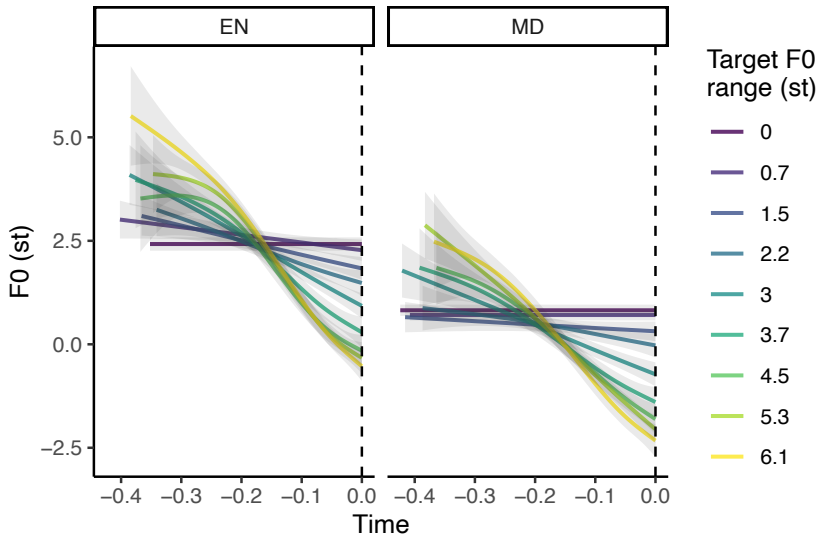


Figure 4.2: Smoothed F0 trajectories of the imitated F0 contours, grouped by continuum step. Trajectories were aligned to the end of the syllable, indicated by the dashed line. Each trajectory is a smooth from a generalized additive model fit to the empirical data for visualization. Shaded areas are the 95% confidence intervals of the smooths.

For each imitation, the F0 range was calculated as the difference between the maximum and minimum of the F0 trajectory. The distribution of F0 range values as a function of continuum step (i.e., each imitation stimulus) is shown in Figure 4.3. The distributions of F0 ranges for both English and Mandarin speakers exhibit two "bumps", indicating a potential categorical pattern in imitation of the continuum. However, compared to Zhang et al. (2023: Figure 5), the distribution for native (Mandarin) speakers appears to be less categorical.

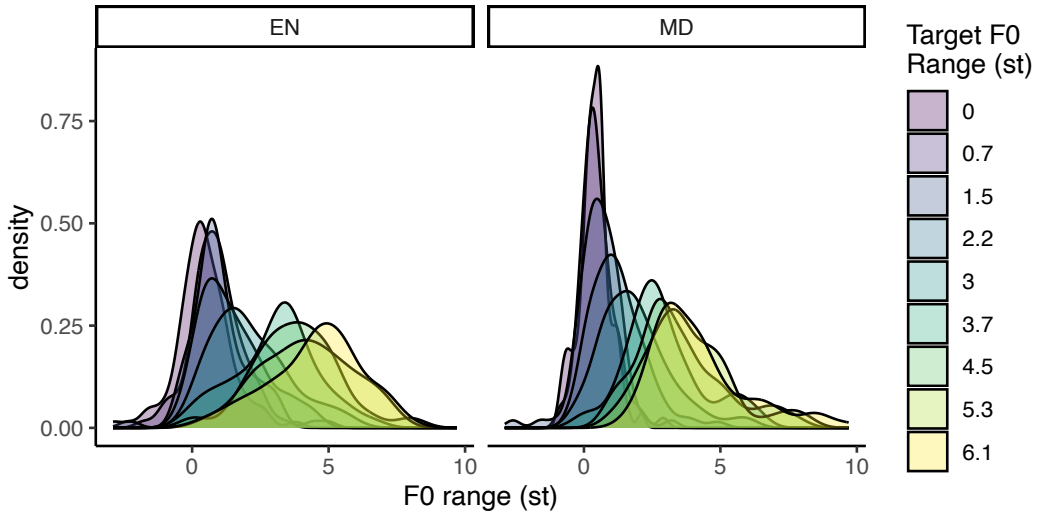


Figure 4.3: Density distributions of the imitated F0 range by English and Mandarin speakers. Colors indicate levels of the target stimuli (i.e., continuum steps).

We consider two statistical models which could account for the distribution of the imitated F0 ranges in Figure 4.3. The two models correspond to the 'unimodal' and 'bimodal' models from Zhang et al. (2023). A similar approach was taken by Massaro & Cohen (1983) in modeling rating responses on a continuous linear scale (also called visual analogue scaling) for stimuli from continua ranging between two categories. To illustrate the two models conceptually, we simulated data that is compatible with our observed data and would be a close fit to the unimodal and bimodal models respectively, shown in Figure 4.4. For the simulation in Figure 4.4 (A), the imitated F0 range increases linearly with the target F0 range, and the variance of the imitated F0 range increases linearly with the target F0 range[5]. This model is designed to reflect a production process in which a talker tries to reproduce the phonetic properties as closely as possible with no role for categories. It will be modeled by a

[5] Note that the variance increasing with the mean is not a requirement of the model but is consistent with our data, as is the case with the variance increasing in the simulation in Figure 4.4(A).

unimodal Gaussian model of which the mean (*mu*) and the standard deviation (*sigma*) can both vary with the target F0 range[6].

For the simulation in Figure 4.4 (B), the imitated F0 range comes from a mixture of two components. One component corresponds to the flat tone category with a smaller mean (for smaller target F0 ranges). The other component corresponds to the falling tone category with a higher mean (for larger target F0 ranges) and a higher variance than the flat tone component[7]. This model is designed to reflect a production process where a talker categorizes the stimulus into one of the two phonological categories and then produces that category. Figure 4.4(C) presents four steps from Figure 4.4 (B) in their own panels, to better illustrate how this model works. In the first panel, productions come mostly from the flat tone category; in the second panel, productions reflect a mix of both categories; by the third and fourth panels, productions mostly come from the falling tone category. In this model, productions come from a mixture of two Gaussian components. The means (*mu1* and *mu2*) and standard deviations (*sigma1* and *sigma2*) of the two components can vary and the rate of mixing is a function of target F0[8].

To determine how well the data are explained by the two models, we fitted both unimodal (i.e., Gaussian) and bi-modal (i.e., two Gaussian mixtures) models to the F0 range data, and compared their goodness of fit. If the bi-modal model provides a significantly better fit, it would suggest that imitation of F0 range is mediated by a phonological contrast, as shown in Figure 4.4 (B). If the unimodal model provides a significantly better fit, it would

[6] This is a 'distributional' version of a standard mixed-effects linear regression model (for mu), where the standard deviation (sigma) is also modelled (Bürkner, 2018; Ciaccio & Veríssimo, 2022; Sonderegger et al., 2023).
[7] Note that this is again not required of the model but is consistent with our data.
[8] This is equivalent to a Gaussian Mixture Model, which has been used to model phonetic categories in previous work (e.g. Franich, 2021; Kirby, 2011; McMurray et al., 2009), but where each GMM parameter is determined by a regression model.

indicate that imitation is less affected by any phonological contrast, as shown in Figure 4.4 (A). However, the data in Figure 4.3 raise the possibility that the imitation data result from a combination of the models in Figure 4.4 (A) and Figure 4.4 (B). In this case, the bi-modal and unimodal models would perform similarly in fitting the data, as each can only explain part of the variability. This would suggest that imitation is affected by phonological contrast as well as within-category phonetic detail. We will consider this possibility using *model averaging*, allowing both models to be weighted together to explain the data (McElreath, 2015).

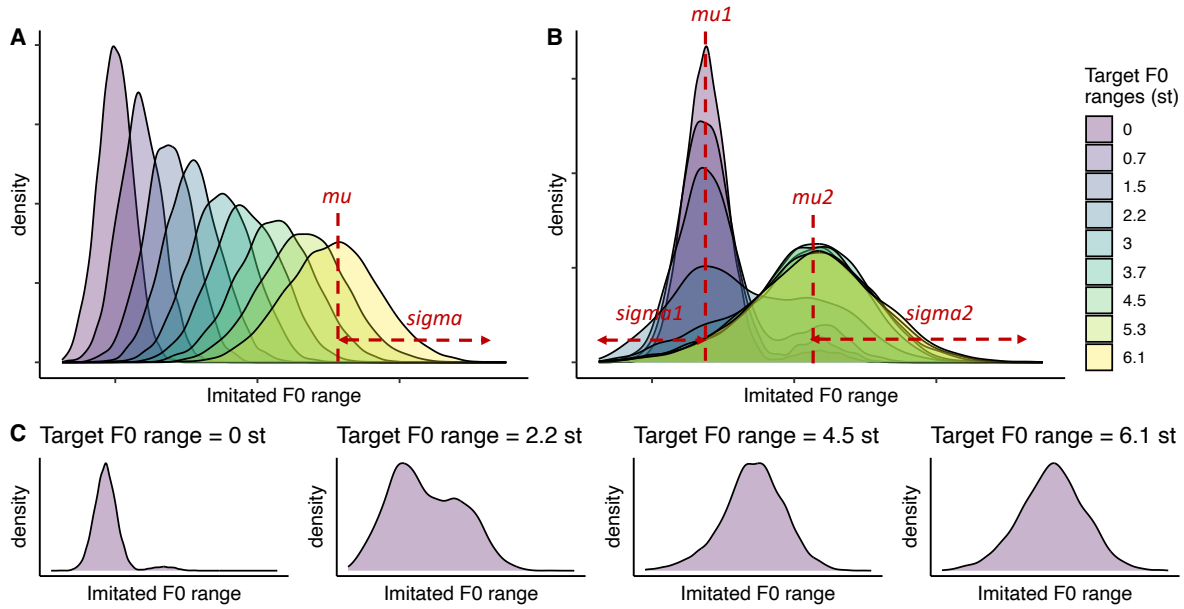


Figure 4.4: Simulated data illustrating the two models of Imitated F0 Range. A: Data simulated from the uni-modal model, with *mu* and *sigma* shown when Target F0 range = 6.1 st. B: Data simulated from the bi-modal model, with *mu* and *sigma* annotated for the two components. C: Four individual steps from the bi-modal model.

The formulas of the simple Gaussian models as specified in *brms* are shown in (1-2). The imitated F0 range is the dependent variable, which follows a Gaussian distribution with mean *mu* and standard deviation *sigma*. The parameter *mu* is modeled by the linear mixed-effects regression in (1). The only fixed effect is target F0 range, and by-participant random intercept

and random slope terms account for interspeaker variability. As seen in the empirical data in Figure 4.3, and captured by the simulation in Figure 4.4 (A), the width of the Gaussian distribution (*sigma*) varies with target F0 range. To capture this, the logarithm of *sigma* is modeled as another dependent variable in (2), again with a fixed effect of target F0 range and by-participant random intercept and random slopes terms[9]. The predictor *target_f0_range* was centered and divided by 2 SD before model fitting.

F0 range ~ 1 + target_f0_range + (1 + target_f0_range | participant) (1)

sigma ~ 1 + target_f0_range + (1 + target_f0_range | participant)
(log link function) (2)

One simple Gaussian model (equations 1-2) is fit to data from each language by drawing 2,000 samples in each of four Markov chains from the posterior distribution over model parameters, with a warm-up period of 1,000 samples per chain. The models are fit using uninformative (flat) priors for fixed-effect terms, and weakly informative regularizing priors for random-effect terms: a half Student's t-distribution with 3 degrees and a brms-default scale of 2.5 for standard deviations, and a LKJ prior with $\eta=1$ for correlations, in order to give lower prior probability to perfect correlations (Vasishth et al., 2018). In all models presented here (Section 4.4.3), parameters are initialized at zero for every chain[10].

The formulas of the bi-modal (i.e., Gaussian mixture) models are (3-8). Conceptually, this model assumes two phonological categories, which are interpolated between as target F0

[9] Note that this is different from the in the simple Gaussian model in Zhang et al. (2023). According to the data distribution, the *mu* and *sigma* in Zhang et al. (2023) were not allowed to vary with target F0 range.
[10] This step was important for model identifiability across chains for the mixture models, and was done for the simple Gaussian models for consistency.

range is increased; the exact form of the categories and 'slope' of interpolation can vary by participant. Similar to the simple Gaussian model, the imitated F0 range is the dependent variable, whose distribution is a mixture of two Gaussians. The means of these Gaussian components, denoted in *brms* as *mu1* and *mu2*, are allowed to vary by participant (by-participant random intercept), with no fixed effects (4-5). Similarly, the standard deviations of the two Gaussian components, *sigma1* and *sigma2*, are allowed to vary by participant, with no fixed effect (6-7). (So that only positive standard deviations can be predicted, it is again the log of standard deviations which are modeled.) The '*theta1'* parameter, which represents the "mixture probability", is the weight of the flat tone category (the first component) in the mixture; thus, (1-*theta1*) is the weight of the falling tone category. Increasing the target F0 range in the stimulus is expected to decrease the probability of the imitation belonging to the first Gaussian component. Therefore, the target F0 range is used as a fixed effect for *theta1*, with by-participant random intercepts and random slopes to allow for variability among participants (8). Because *theta1* is a probability, a logit link is used in the regression (8) so that only *theta1* values between 0 and 1 can be predicted.

Unlike previous models, some fixed effects have informative priors, to enable the model to converge on a unique solution (called "identifiability"), as is often necessary for non-linear Bayesian models (Bürkner, 2018). As the simple Gaussian models, 2,000 samples are drawn in each Markov chain to ft the data and 1,000 of them are used for warm-up period. To ensure the first component has a lower f0 range than the second component, priors are set to normal(0.5, 0.3) for *mu1* and normal(3.0, 0.3) for *mu2*. A prior of normal(-15, 3) is used for the effect of target F0 range, which indicates a negative effect (corresponding to the first

component having a lower mean)[11]. The model's priors are otherwise the same as the simple Gaussian model: uninformative (for fixed effects) or weakly informative regularizing (for random effects).

$$\text{F0\_range} \sim 1, \tag{3}$$

$$\text{mu1} \sim 1 + (1 \mid \text{participant}), \tag{4}$$

$$\text{mu2} \sim 1 + (1 \mid \text{participant}), \tag{5}$$

$$\text{sigma1} \sim 1 + (1 \mid \text{participant}), \quad \text{(log link)} \tag{6}$$

$$\text{sigma2} \sim 1 + (1 \mid \text{participant}), \quad \text{(log link)} \tag{7}$$

$$\text{theta1} \sim 1 + \text{target\_f0\_range} + (1 + \text{target\_f0\_range} \mid \text{participant}), \quad \text{(logit link)} \tag{8}$$

The performance of the simple Gaussian and the Gaussian-mixture models are compared using Pareto smoothed importance sampling leave-one-out cross-validation (PSIS-LOO CV), using the *loo* package (Vehtari et al., 2021). The full model summaries of the Gaussian models in this section and the following section can be accessed in the OSF repository. Model comparison results are summarized in Table 4.3. The difference of the expected log pointwise predictive density (ELPD) between the simple and the mixture models suggests that the mixture model is a slightly better fit for the Mandarin group's imitation whereas the unimodal model is a slightly better fit for the English group. However, for both groups the difference in goodness of fit between the two models is small, as both comparisons

[11] The mean value of this prior, -15, was approximated using a preliminary model with a single fixed effect. The standard deviation of 3 makes the prior only weakly informative about the magnitude of the slope, but informative about its direction.

show relatively large standard errors, with the 2SE range around the ELPD difference including 0, especially in the English comparison.

Table 4.3: Model comparison results (*ELPD*: log pointwise predictive density. *SE*: standard error)

Fits for Mandarin group		Fits for English group	
Model	ELPD difference (SE)	Model	ELPD difference (SE)
Mixture	**0 (0)**	**Unimodal**	**0 (0)**
Unimodal	-26.1 (22.7)	Mixture	-11 (20.7)

Results in Table 4.3 indicate that the imitation of both groups is neither purely unimodal nor purely bi-modal but is a mix of both, neither the pure unimodal nor the pure bi-modal one significantly outperforms the other. It could be that the imitation has both categorical characteristics and a linear trend within each category, which aligns with the visualization in Figure 4.3. To further examine this possibility, we calculate the optimal combination of the two models that best describes the data, so-called *model averaging* (McElreath, 2015: Sec. 5.8). We determine the weights of each model using the stacking method (Yao et al., 2018), implemented in the *loo* package, using the *loo_model_weight()* function (Vehtari et al., 2021). The weights are:

- Mandarin: uni-modal model 0.41, bi-modal model 0.59
- English: uni-modal model 0.52, bi-modal model 0.48.

That is, the Mandarin imitation data are best described by giving 41% weight to the unimodal model and 59% weight to the bi-modal model, and similarly for the English

imitation data. For both English and Mandarin speakers, both the unimodal and the bi-modal models are necessary to best describe the data. This supports the idea that both pre-existing contrasts and phonetic details are used in imitation. In addition, Mandarin imitation is more categorical than the English imitation, as the best model for the Mandarin data assigns more weight to the bi-modal part than in the English data.

4.4.4 Unplanned analyses

Results in Section 4.4.3 showed that both Mandarin speakers and English speakers imitated the linearly-varying F0 ranges into two categories while preserving some within-category variation. This finding is noteworthy because the English speakers were naïve to Mandarin and lacked the flat vs. falling lexical tonal contrast in their native language. These results address our main research question, showing that phonetic imitation is not solely determined by *phonological* contrasts, which only the Mandarin speakers had in this study. We also found that English speakers' imitation was less categorical than that of Mandarin speakers. It is of interest to investigate any other differences between native and non-native speakers in the categorical aspect of their imitations, to understand what kind of categories underlie their behavior.

Since English speakers do not have tonal categories, their categories must be at other levels. This conjecture implies that lexical categories should vary in other dimensions that cue the contrast, such as duration, while non-lexical categories should not. Although duration is not the primary cue used to contrast tones in Mandarin, the falling tone is shorter than the flat tone (Xu 2005), and duration has been found to influence the categorization of the flat vs. falling tonal contrast as a secondary cue (Zhang et al., 2023). As discussed earlier with respect

to Figure 4.2, imitations of steps 6-9 are shorter than imitations of steps 1-4 for Mandarin but not English speakers. Figure 4.5 shows the empirical relationship between F0 range in the continuum and the imitated duration for both groups. F0 step seems to have a negative effect on duration for Mandarin participants but no effect for English participants. To test the statistical significance of these patterns, we fit mixed-effect Bayesian linear regression model of the imitated duration; with fixed effects of *target f0 range*, participant *Group* and their interaction; and by-participant random intercepts and random slopes matching the fixed effects. The predictor *Group* was dummy coded with English as the reference, and the predictor *target f0 range* was centered and divided by 2 SD before model fitting. Model fitting used 2000 draws (1000 warm-up) from each of four chains, with the same priors as used for the simple Gaussian model.

The model's fixed effects are presented in Table 4.4. Target F0 range does not credibly affect duration for the English group, which serves as the baseline in the dummy coding (estimate = 0, 95% CI = [-0.01, 0.01]), consistent with the flat line for English participants in Figure 4.5. The participant group credibly interacts with the effect of target F0 range (estimate = -0.03, 95% CI = [-0.05, -0.01]), indicating that target F0 range has a more negative effect on duration for the Mandarin group than for the English group.

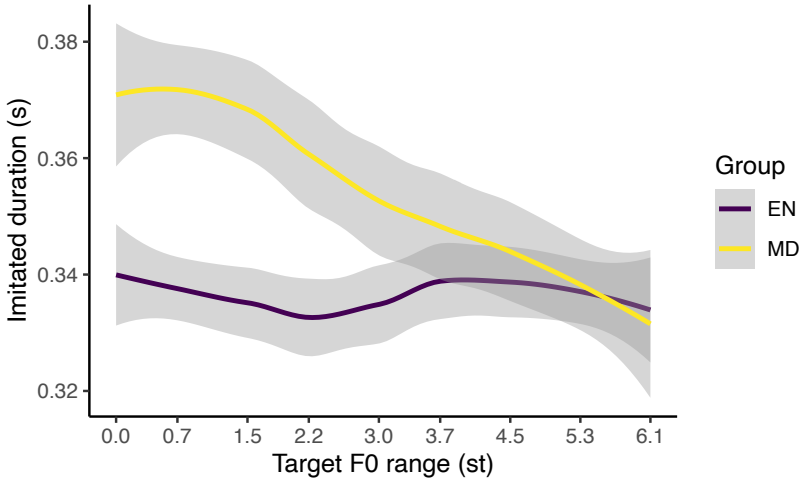


Figure 4.5: Imitated duration as a function of target F0 range, for both groups of participants. Each curve is a smooth from a generalized additive model fit to the empirical data for visualization. Shaded areas are the 95% confidence intervals of the smooths.

Table 4.4: Results of the mixed-effect Bayesian linear regression model for the imitated durations

	Estimate	Std.Error	l-95% CI	u-95% CI	PD
(Intercept)	0.34	0.01	0.32	0.35	**100%**
target f0 range	0.00	0.00	-0.01	0.01	63%
Group (MD-EN)	0.02	0.02	-0.01	0.05	87%
target f0 range* Group (MD-EN)	-0.03	0.01	-0.05	-0.01	**99%**

The results suggest that the representation of the categories differs qualitatively between English and Mandarin groups. For Mandarin participants, the falling category is a lexical tone category, which is associated with a shorter duration compared to the flat tone category. In contrast, the falling category imitated by English participants does not resemble this lexical tone category. As discussed further in Section 4.5, the falling category guiding the imitation of English participants may be a weaker one, such as the psychophysical distinction between

flat and falling F0 shapes. On the other hand, the lexical tone category of Mandarin participants is stronger, and can influence the reproduction of multiple phonetic dimensions of the tonal category. This difference in category strength may explain the greater variation in F0 range imitation for English participants compared to Mandarin participants observed in Section 4.4.3, as well as the reduced variation along other lexical tone-related phonetic dimensions.

Within the falling tone category, native Mandarin speakers may interpret larger F0 ranges as hyper-articulation or emphasized tone production (De Jong, 1995; Lindblom, 1990). This hyper-articulation of the tone may be accompanied by other acoustic changes, such as expanded vowel space. Since the carrier vowel is the low vowel /a/, if larger falling F0 ranges are interpreted as hyper-articulation, we expect the imitated productions to show a lower jaw position and a more open vocal tract, leading to increased F1 and intensity (De Jong, 1995). We now test whether the expanded F0 range induces hyper-articulation in F1 and intensity. We predict that hyper-articulation will be present in the imitations of Mandarin participants but not in those of of English participants.

Figure 4.6 illustrates the effects of target F0 range on F1 and intensity, averaged over the imitated syllable, for both Mandarin and English groups. Note that to account for anatomical differences, F1 and intensity have been centered within speaker for visualization. For the Mandarin group, both F1 and intensity show a trough around step 4-5 (2.2-3 st, or 30-40 Hz), which corresponds to the categorical boundary observed in the perceptual results in Section 4.3.3. They peak at the first and the last continuum steps, where the tones are clear and hyper-articulated. However, there is less change in F1 or Intensity as a function of target F0 range in the English group. To examine the effects of target F0 range on F1 and Intensity,

because the empirical relationships looked nonlinear, we used GAMM models, which allow for analysis of data with non-linear relationships without making any assumption about the functional form of the relationship (as in e.g. growth curve models: Wood, 2017). In these models, F1 or intensity was the dependent variable; participant *Group* was a parametric term (dummy coded with English as the reference level); and a smooth of *Target F0 range* was included for each participant *Group*, along with a by-participant random smooth of Target F0 range, to reliably estimate the effect for each *Group*. For English participants, the target F0 range was not significantly different from a flat line for either F1 (edf = 2.5, p = 0.2) or intensity (edf = 1, p = 0.62). In contrast, for Mandarin participants both the smooth for F1(edf = 3.5, p < 0.001) and intensity (edf = 2.2, p < 0.003) were significantly different from a flat line, with the fitted smooths resembling the MD curves in Figure 4.6. Thus, target F0 range influenced the F1 and intensity of the imitations in different ways, for the two groups. Full model results can be found in the Appendix 4.

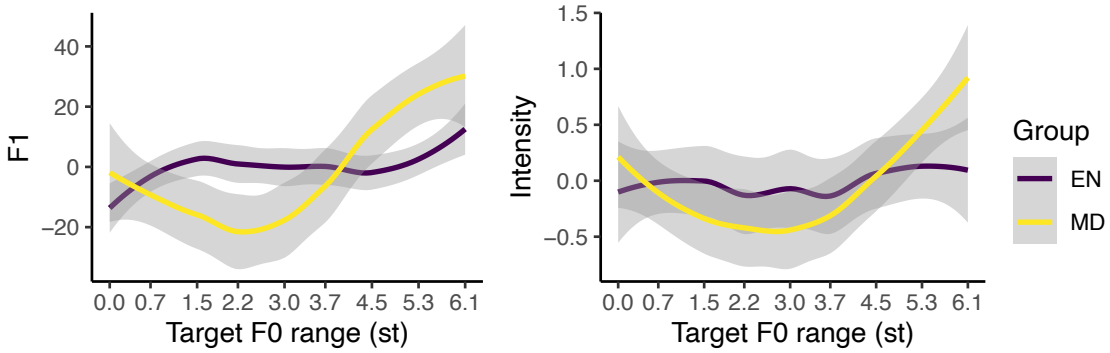


Figure 4.6: Imitated F1 and intensity (both centered) as a function of target F0 range, for the two groups of participants. Each curve is a smooth from a generalized additive model fit to the empirical data for visualization. Shaded areas are the 95% confidence intervals of the smooths.

4.5 Discussion

Our previous study (Zhang et al., 2023) investigated how native Mandarin speakers imitated varying F0 contours between the Mandarin flat and falling tones. It was found that phonetic imitation was mediated by the native tonal contrast, which appeared to be related to nonlinearity in perception. An earlier study by Wang (1976) had proposed that perception of Mandarin tonal contrasts involves both a linguistic boundary and a psychophysical boundary. The linguistic boundary was observed only in speakers with tonal phonology, whereas the psychophysical boundary was present in all speakers, regardless of their tonal phonology. Building on this idea, the current study examined whether nonlinearity in phonetic imitation is influenced by the linguistic boundary or by the psychophysical boundary. To address this question, we compared how native Mandarin speakers and naïve English speakers imitated F0 contours ranging between Mandarin flat and falling tones. If both groups show nonlinearity in their imitation, it would suggest that phonetic imitation is mediated by the psychophysical boundary, which should be present in both groups. However, if only the Mandarin speakers' imitation is non-linear while the English speakers' imitation is linear, it would suggest that phonetic imitation is primarily influenced by the linguistic boundary, which is specific to speakers with the relevant tonal phonology.

To assess the nature of the distribution of imitations by Mandarin and English speakers, we compared the performance of uni-modal and bi-modal models on each groups' imitation data. Both models performed similarly for both groups, although the bi-modal model was a somewhat better fit for the Mandarin data while the uni-modal model was a slightly better fit for the English data. Using Bayesian model averaging, we found that a combination of the uni-modal and bi-modal models best fits the imitation data, with the Mandarin group's

imitation being more heavily influenced by the categorical (bi-modal) component compared to the English group. These findings show that both pre-existing categories and within-category phonetic details are involved in phonetic imitation. Crucially, regarding the main research question, the results support the first scenario from Section 4.1.4: both Mandarin and English speakers showed one flat category and one falling category overall in their imitation distributions. To our knowledge, this is the first study to find that phonemic imitation is influenced by both stronger, higher-level linguistic contrast, and weaker, lower-level psychophysical categories (though see Section 4.5.1 for discussion of mediation by prosodic categories).

Post-hoc analyses revealed other differences in imitation by Mandarin and English speakers, which may provide further evidence for the different types of categories involved in imitation by the two groups. First, Mandarin speakers' imitations become shorter as the falling contours get steeper, whereas the durations of English speakers' imitations are not affected by the F0 shape. Second, signs of hyper-articulation, characterized by larger F1 and intensity values for larger F0 ranges, are observed in Mandarin speakers' imitations but not in those of English speakers. Finally, categorization results of the same tonal continuum (from another group of English and Mandarin speakers) show that English speakers' categorization is less categorical than that of native Mandarin speakers, and that they identify more stimuli as falling category than Mandarin speakers. These results will be further discussed in the following sections.

4.5.1 Implications for the mechanism of phonetic imitation

Results of the model comparison in the present study show that both the uni-modal or the bi-modal models are necessary to explain variations in imitation distribution for both groups. For both groups, there are both categorical patterns and continuous variations within those categories. This dual-pattern structure suggests that participants use a combination of discrete and continuous information when imitating the target F0s in the stimuli. Imitators interpret the F0 contours as a flat or falling category, and keep track of how flat or how steep the contour is. These findings are consistent with studies that found phonological mediation effects in imitation (Chistovich et al., 1966; Flege & Eefting, 1988; Kent, 1973; Mitterer & Ernestus, 2008; Repp & Williams, 1985), studies that observed linear imitation patterns (Dilley, 2010; Michalsky, 2015; Podlipskỳ & Simácková, 2015), and those that found tendencies of both patterns (Kim & Clayards, 2019; Nielsen, 2011; Schertz et al., 2023; Schertz & Johnson, 2022; Zhang et al., 2023), highlighting the coexistence of phonological and phonetic information in the imitation process.

Although naïve English speakers did not have the lexical tone phonology, they still interpreted the continuum as consisting of an overall flat category and a falling category. However, evidence was found indicating that the falling category differed in essence between the two groups of participants' imitation. Although all stimuli had the same length, Mandarin speakers' imitation exhibited a shorter average duration for the falling category as the F0 range in the stimulus increased. This is most likely because the falling tone is shorter than the flat tone in Mandarin (Xu, 2005), and native speakers reproduced the F0 contours guided by Mandarin tone phonology, resulting in the shorter duration being reflected in their imitation of the F0 falling contours. However, for English speakers, the imitated duration was not

affected by the F0 range in the stimulus. The lack of the representation of a shorter falling tone led to their duration imitation being unaffected. Furthermore, we found that only native speakers, but not English speakers, imitated larger falling ranges accompanied by higher F1 and higher vowel intensity – both of which are indicators of hyper-articulation given the carrying vowel /a/ (De Jong, 1995; Lindblom, 1990). This observation indicated that native speakers interpreted larger falling ranges (compared to a medium amount of F0 fall) as hyper-articulated falling tones, and then the hyper-articulation was reproduced holistically across multiple dimensions. However, naïve English speakers, who lacked such a hyper-articulation interpretation, were only imitating the F0 dimension.

The discrepancy revealed differences in how Mandarin and English speakers interpreted the F0 falling contours. The differences are unsurprising, given that naïve English speakers do not have the same Mandarin falling tone category. However, it is interesting to ask what category is involved in the English speakers imitation of the falling F0 contours. According to our hypothesis, the falling category involved in English speakers' imitation may be a psychophysical falling category, as suggested by Wang et al (1976). This category can be characterized as a 'non-flat' F0 shape in a psychophysical sense, which is consistent with our categorization result that the categorization boundary position of the English group is closer to the flat end of the continuum than that of the Mandarin group, indicating that more stimuli are perceived as instances of the falling category by English speakers than Mandarin speakers. Additionally, the psychophysical falling category is a lower-level category which should have less relation to other dimensions such as vowel space, aligning again with our results that the duration, intensity or vowel formants in the English speakers' reproduction were not affected.

As reviewed in Section 4.1.2, intonational categories (such as statement and flat-pitch intonations) might have been involved in the imitation of the flat-falling tonal contrast. However, assimilation results have revealed that the flat and falling tones in Mandarin were not a simple or direct mapping onto two intonational categories (So and Best, 2008, 2011, 2014). Additionally, discrimination results of the Mandarin flat-falling tonal contrast by naïve English speakers have been mixed, with some studies reporting that naïve English speakers had difficulty discriminating between them (Hao, 2018; Shen, 1989; So & Best, 2010) while others did not (Bent, 2005). Bent (2005) noted that when Mandarin tones are embedded in more varied tonal contexts, English speakers' identification performance varies, hence she speculated that the relevant English intonational categories also vary. Furthermore, Steffman et al., (2024) found that there may not exist any flat-pitch intonation in English. In short, it remains unclear whether naïve English speakers perceive the Mandarin flat tone and falling tone as a single intonational category or discrete categories. It could be that studies so far have not used the correct intonational categories, or it could be that the flat tone doesn't map neatly onto a natural category for English. While we cannot rule out the possibility of intonational categories, the existence of psychophysical categories remains a possible explanation.

4.5.2 Pitch range difference between the imitation of Mandarin and naïve English speakers

Previous results on Mandarin tone imitation by naïve English speakers showed that they imitated the canonical tones with a compressed pitch range (Bent, 2005). The two groups in the present study do not show differing pitch ranges in their imitation. However, we find that English speakers use higher F0 level than native speakers, as compared to their native F0 levels. There are two potential reasons for this discrepancy. First, this study uses a continuum

between the flat and the falling tone whereas Bent (2005) employed canonical productions of the four tones. The variation of stimuli in a block could affect participants' attention to cues (Holt and Lotto, 2006). Being exposed to both clear and intermediate cases of F0 range, as in the case of the flat-falling continuum, and being exposed to four distinct canonical tones may elicit different attention to the cues. Second, for F0 normalization, Bent (2005) used T values (Ladd et al.,1985), which normalize each speaker's raw F0s by dividing the speaker's pitch range into 6 values (0 to 5) in order to correspond to the labels used by Chao (1965). In contrast, the analysis in the current study uses semitones, with the reference F0 for each speaker being the mean F0 from their native speech. In other words, the current study compared the F0s relative to the English speakers' mean native production, whereas Bent (2005) analyzed the F0 values relative to the range of their L2 productions, making direct comparisons difficult. Results in the present study indicate that native and nonnative speakers interpret the F0 levels differently relative to their own language, with nonnative speakers potentially adjusting their F0 level upward when attempting to match the target tones.

4.5.3 Imitation as a paradigm

Categorization experiments of the same set of stimuli were carried out for both English and Mandarin speakers. Results showed that English speakers' categorization was less categorical than that of native Mandarin speakers, and that the categorization boundary position was at a smaller F0 range that that of Mandarin speakers. Other than the degree of categoricalness and the categorization position, categorization experiments provide limited information about the specific characteristics of each category in the contrast for Mandarin and English speakers. However, phonetic imitation of the continuum offered more detailed insights for each category.

The imitation results reveal that both native and non-native speakers remain sensitive to within category differences while making categorical distinctions. For both groups, the imitated F0 range increases with the target F0 range within each category. Such within-category phonetic imitation, which replicated Zhang et al. (2023), has also been observed in the imitation of stop VOT: imitation of both shortened, canonical and lengthened VOT has been observed (Nielsen, 2011; Schertz and Johnson, 2022; Schertz et al., 2023). Other studies also found evidence of within-category sensitivity in the perception of Mandarin tones. Through a visual-word eye-tracking paradigm, Qin et al. (2019) found that for both Mandarin speakers and non-native Mandarin learners, word activation was affected by within-category tonal variation. In an event-related potential (ERP) experiment using the oddball paradigm, Gao et al. (2019) investigated the perception of the Mandarin falling tone. They presented deviant stimuli that were either within- or across- category in relation to the standard falling tone. If the perception was entirely categorical, within-category deviants would be expected to elicit lower P3 amplitudes, as P3 was often regarded to index categorization. However, they found that within-category deviants induced larger P3 amplitudes than across-category deviants. Based on these findings, Gao et al. (2019) concluded that listeners remain sensitive to fine-grained phonetic variations in speech. This is consistent with a growing body of studies that highlight the gradient nature of speech perception in contrast to category effects (see McMurray (2022), for a review). Whether a more categorical/phonological or more gradient/phonetic pattern is observed depends greatly on the experimental paradigm and task demands (Strange, 2011). In the context of our study, our results support the view that phonetic imitation involves both phonetic (continuous) and phonological (categorical)

processing, as evidenced by the presence of both categorical and within-category effects in our imitation data.

These imitation results of Mandarin tones for naïve English speakers reflect some interesting differences from the perceptual results in previous literature which indicated that English listeners pay less attention to F0 shape than F0 height (Gandour 1983, Francis et al., 2008, Huang, 2004). First, we found that English speakers used an obviously higher-than-average pitch in imitating the flat tone, unlike Mandarin speakers who used a pitch height close to their average level. Second, English speakers were able to reproduce the F0 fall variations within the falling category. These results seem to suggest poor pitch height matching but good pitch contour matching, contrary to previous perception results. One possible reason for this difference is that the two paradigms reflect different stages of processing. In the imitation paradigm, participants are asked to imitate the stimulus as faithfully as possible immediately after the stimulus was played. In contrast, the perception paradigm involves extra cognitive and linguistic processing, as participants need to interpret and respond to questions related to the stimulus. Compared to answering questions about what participants hear, it appears that imitation is more direct in reflecting what participants hear. The observation that naïve English speakers performed better in imitating than identifying/discriminating F0 contours (Bent 2005; Hao 2012; Hao and de Jong, 2016) suggests that naïve participants can perceive the F0 shape variations but are not good at categorizing them. It is possible that the weaker categorization ability leads to poorer performance in answering questions related to the stimulus in the perception paradigm. This discrepancy between imitation and perception performance highlights the importance of considering different stages and processes involved in speech perception tasks.

4.5.4 Limitations and future directions

There are also limitations and questions that remain open for exploration in future studies. The categorization result of the English speakers differs from that reported in Peng et al. (2010), where naïve German speakers responded to extreme steps with a rate of 0 or 100%, but is similar to the pattern of naïve English speakers in Shen and Froud (2016). We also note that English participants showed qualitatively different types of behaviour, with 6 participants responding randomly over the whole continuum, while 9 consistently categorized the endpoints. Such variation, which is presumably present in previous studies of non-native speakers as well (e.g. in Shen and Froud (2016)), complicates comparison. Plots for individual speakers are included in supplemental materials in the OSF repository. This discrepancy is likely due to varying degrees of training received by these naïve speakers before the main session of the experiment. In the current experiment, the training may have been less effective compared to the previous two, as data were collected remotely without real-time instruction by an experimenter. While participants had the option to play the example audio of the tones an unlimited number of times, English participants were only provided with two examples of each tone.

Comparing the results of the native Mandarin speakers in Zhang et al. (2023) to the current study, based on the data visualization, the participants in the current study seemed to show a less categorical imitation pattern. This indicates that the mediation effects of tone categories were stronger in Zhang et al. (2023) than in the current study, highlighting the flexibility of the mediation effects, which may result from the following factors. First, Zhang et al. (2023) employed a semantically neutral carrier (meaning 'the sound of') in the imitation whereas this study used the isolated form. Previous research (e.g., Francis et al., 2003) has

shown that context can enhance the categorization effect in the perception of Mandarin tones. The difference in gradiency of imitation observed between Zhang et al. (2023) and the current study suggests that this context effect may extend from perception to production in an imitation task. The lack of a context may strengthen the auditory processing mode hence bringing about more gradient imitation performance. Additionally, using the isolated form versus the carrier form resulted in different durations of the target syllable (longer in this study), as well as the overall F0 height of the continuum (lower in this study), as they were calculated from the natural productions of the trisyllabic phrases and the isolated words, respectively. Longer syllables may provide more cognitive resources which may facilitate more subtle perception, and longer productions also allow participants the opportunity for finer control of production. Second, there are 9 steps in the present study whereas there were 7 in Zhang et al. (2023). As the maximum fall ranges across the two experiments are similar, this leads to a smaller step size in this study. The more fine-grained steps in this study might have given us more opportunity to observe gradient behaviour. Future work could examine to what extent each of these factors may influence imitation gradiency.

Although we interpret the categories in the English speakers' imitation as flat and falling psychophysical categories, we cannot entirely rule out the possibility that intonational categories play a role. Future research could explore intonational contrast in English, and investigate how they might mediate the imitation of F0 contours in more nuanced ways.

4.5.5 Conclusion

This study investigates the mechanism of phonetic imitation by comparing the categorization and imitation of the Mandarin flat-falling tonal continuum between native Mandarin speakers

and English speakers who lack tone experience. In their imitation, both groups showed both flat vs. falling categorical distinctions as well as within-category detail in the degree of F0 fall. These results support the view that phonetic imitation involves both phonological and phonetic information.

The presence of two categories in the imitation of English speakers', who do not have tonal categories, indicates that phonetic imitation is not merely mediated by linguistic categories. Post-hoc analyses revealed notable differences between the Mandarin speakers' categories and the English speakers' categories. Imitation was more guided by the flat vs. falling categories for the Mandarin participants than for the English participants, and only Mandarin speakers imitated larger F0 ranges with hyper-articulation and category-specific duration. Additionally, in the categorization results, English speakers showed weaker categorization effects and identified stimuli as the falling category more frequently than Mandarin listeners. These discrepancies align with the idea that English speakers' imitation of the F0 ranges is mediated by low-level psychoacoustic categories, in contrast with the linguistic categories of Mandarin speakers.

References

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice*. Cengage Learning.

Bent, T. (2005). *Perception and production of non-native prosodic categories* [Ph.D., Northwestern University]. https://www.proquest.com/docview/305432272/abstract/5435C43C19947C0PQ/1

Boersma, P. (2021). Praat: Doing phonetics by computer [version 6.1.38]. *Http://Www. Praat. Org/*.

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411.

Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, *128*(1), 456–465. https://doi.org/10.1121/1.3445785

Chang, D., Hedberg, N., & Wang, Y. (2016). Effects of musical and linguistic experience on categorization of lexical and melodic tones. *The Journal of the Acoustical Society of America*, *139*(5), 2432–2447. https://doi.org/10.1121/1.4947497

Chistovich, L., Fant, G., de Serpa-Leitao, A., & Tjernlund, P. (1966). Mimicking of synthetic vowels. *Quarterly Progress and Status Report, Speech Transmission Lab, Royal Institute of Technology, Stockholm*, *1*(2), 1–18.

Ciaccio, L. A., & Veríssimo, J. (2022). Investigating variability in morphological processing with Bayesian distributional models. *Psychonomic Bulletin & Review*, *29*(6), 2264–2274. https://doi.org/10.3758/s13423-022-02109-w

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809. https://doi.org/10.1016/j.cognition.2008.04.004

Cole, J., Steffman, J., Shattuck-Hufnagel, S., & Tilsen, S. (2023). Hierarchical distinctions in the production and perception of nuclear tunes in American English. *Laboratory Phonology*, *14*(1). https://doi.org/10.16995/labphon.9437

De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, *97*(1), 491–504.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

De Looze, C., & Hirst, D. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. *Speech Prosody*, 4.

Dilley, L. C. (2010). Pitch Range Variation in English Tonal Contrasts: Continuous or Categorical? *Phonetica*, *67*(1–2), 63–81. https://doi.org/10.1159/000319379

Flege, J. E., & Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: Evidence for phonetic category formation. *The Journal of the Acoustical Society of America*, *83*(2), 729–740. https://doi.org/10.1121/1.396115

Francis, A. L., Ciocca, V., Ma, L., & Fenn, K. (2008). Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. *Journal of Phonetics*, *36*(2), 268–294. https://doi.org/10.1016/j.wocn.2007.06.005

Franich, K. (2021). Metrical prominence asymmetries in Medumba, a Grassfields Bantu language. *Language*, *97*(2), 365–402. https://doi.org/10.1353/lan.2021.0021

Gandour, J. (1983). Tone perception in Far Eastern languages. *Journal of Phonetics*, *11*(2), 149–175. https://doi.org/10.1016/S0095-4470(19)30813-7

Gao, Y. A., Toscano, J. C., Shih, C., & Tanner, D. (2019). Reassessing the electrophysiological evidence for categorical perception of Mandarin lexical tone: ERP evidence from native and naïve non-native Mandarin listeners. *Attention, Perception, & Psychophysics*, *81*(2), 543–557. https://doi.org/10.3758/s13414-018-1614-8

Hallé, P. A., Chang, Y.-C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*(3), 395–421. https://doi.org/10.1016/S0095-4470(03)00016-0

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

Hao, Y.-C. (2018). Second Language Perception of Mandarin Vowels and Tones. *Language and Speech*, *61*(1), 135–152. https://doi.org/10.1177/0023830917717759

Hao, Y.-C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics*, *54*, 151–168.

Huang, T. (2004). *Language specificity in auditory perception of Chinese tones* [Ph.D., The Ohio State University]. https://www.proquest.com/docview/305139053/abstract/16C40D6C6D7B4802PQ/1

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, *97*(1), 553–562.

Kent, R. D. (1973). The imitation of synthetic vowels and some implications for speech memory. *Phonetica*, *28*(1), 1–25.

Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, *34*(6), 769–786. https://doi.org/10.1080/23273798.2019.1582787

Kirby, J. P. (2011). Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association*, *41*(3), 381–392. https://doi.org/10.1017/S0025100311000181

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech production and speech modelling* (pp. 403–439). Springer.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, *2*(1), 15–35. https://doi.org/10.1016/0167-6393(83)90061-4

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech 2017*, *2017*, 498–502.

McElreath, R. (2015). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

McMurray, B. (2022). *The Myth of Categorical Perception*.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1609–1631. https://doi.org/10.1037/a0011747

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 65–91. https://doi.org/10.1016/j.jml.2008.07.002

Michalsky, J. (2015). Pitch scaling as a perceptual cue for questions in German. *Sixteenth Annual Conference of the International Speech Communication Association*.

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. https://doi.org/10.1016/j.cognition.2008.08.002

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, *10*(11), 591–613. https://doi.org/10.1111/lnc3.12207

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142. https://doi.org/10.1016/j.wocn.2010.12.007

Peng, G., Zheng, H.-Y., Gong, T., Yang, R.-X., Kong, J.-P., & Wang, W. S.-Y. (2010). The influence of language experience on categorical perception of pitch contours. *Journal of Phonetics*, *38*(4), 616–624. https://doi.org/10.1016/j.wocn.2010.09.003

Pierrehumbert, J. B., & Steele, S. A. (1989). Categories of tonal alignment in English. *Phonetica*, *46*(4), 181–196.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, *13*(2), 253–260.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*(2), 285–290. https://doi.org/10.3758/BF03213946

Podlipskỳ, V. J., & Simácková, S. (2015). Phonetic imitation is not conditioned by preservation of phonological contrast but by perceptual salience. *ICPhS*.

Qin, Z., Tremblay, A., & Zhang, J. (2019). Influence of within-category tonal information in the recognition of Mandarin-Chinese words by native and non-native listeners: An eye-tracking study. *Journal of Phonetics*, *73*, 144–157.

Repp, B. H., & Williams, D. R. (1985). Categorical trends in vowel imitation: Preliminary observations from a replication experiment. *Speech Communication*, *4*(1–3), 105–120.

Schertz, J., Adil, F., & Kravchuk, A. (2023). Underpinnings of explicit phonetic imitation: Perception, production, and variability. *Glossa Psycholinguistics*, *2*(1). https://doi.org/10.5070/G601123

Schertz, J., & Johnson, E. K. (2022). Voice Onset Time Imitation in Teens Versus Adults. *Journal of Speech, Language, and Hearing Research*, *65*(5), 1839–1850. https://doi.org/10.1044/2022_JSLHR-21-00460

Shen, G., & Froud, K. (2016). Categorical perception of lexical tones by English learners of Mandarin Chinese. *The Journal of the Acoustical Society of America*, *140*(6), 4396–4403. https://doi.org/10.1121/1.4971765

Shen, G., & Froud, K. (2019). Electrophysiological correlates of categorical perception of lexical tones by English learners of Mandarin Chinese: An ERP study. *Bilingualism: Language and Cognition*, *22*(2), 253–265. https://doi.org/10.1017/S136672891800038X

Shen, X. S. (1989). Toward a register approach in teaching Mandarin tones. *Journal of the Chinese Language Teachers Association*, *24*(3), 27–47.

So, C. K., & Best, C. T. (2008). Do English speakers assimilate Mandarin tones to English prosodic categories? *INTERSPEECH*, 1120.

So, C. K., & Best, C. T. (2010). Cross-language Perception of Non-native Tonal Contrasts: Effects of Native Phonological and Phonetic Influences. *Language and Speech*, *53*(2), 273–293. https://doi.org/10.1177/0023830909357156

So, C. K., & Best, C. T. (2011). Categorizing Mandarin tones into listeners' native prosodic categories: The role of phonetic properties. *Poznań Studies in Contemporary Linguistics*, *47*. https://doi.org/10.2478/psicl-2011-0011

So, C. K., & Best, C. T. (2014). PHONETIC INFLUENCES ON ENGLISH AND FRENCH LISTENERS' ASSIMILATION OF MANDARIN TONES TO NATIVE PROSODIC CATEGORIES. *Studies in Second Language Acquisition*, *36*(2), 195–221. https://doi.org/10.1017/S0272263114000047

Sonderegger, M., Stuart-Smith, J., & Mielke, J. (2023). *How variable are English sibilants?* https://doi.org/10.17605/OSF.IO/XUBQM

Strange, W. (2011). Automatic selective perception (ASP) of first and second language speech: A working model. *Journal of Phonetics*, *39*(4), 456–466. https://doi.org/10.1016/j.wocn.2010.09.001

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161. https://doi.org/10.1016/j.wocn.2018.07.008

Vehtari, A., Gelman, A., Gabry, J., & Yao, Y. (2021). Package 'loo.' *Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. https://cran.r-hub.io/web/packages/loo/loo.pdf

Wagner, M. (2021). *Prosody lab Experimenter* [JavaScript]. Prosodylab. https://github.com/prosodylab/prosodylabExperimenter (Original work published 2020)

Wang, W. S. Y. (1976). Language Change. *Annals of the New York Academy of Sciences*.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*(3–4), 220–251. https://doi.org/10.1016/j.specom.2005.02.014

Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America*, *120*(2), 1063–1074. https://doi.org/10.1121/1.2213572

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, *13*(3). https://doi.org/10.1214/17-BA1091

Zhang, W., Clayards, M., & Torreira, F. (2023). Phonological mediation effects in imitation of the Mandarin flat-falling tonal continua. *Journal of Phonetics*, *101*, 101277. https://doi.org/10.1016/j.wocn.2023.101277

Appendix 4.

Table A1: Summary of the GAM model for imitated F1 (edf: estimated degrees of freedom, Ref.df: reference degrees of freedom, the same in Table A2).

	Estimate	SE	p
(Intercept)	798	21.2	<0.001
LANG (MD-EN)	4.4	30.9	0.89
	edf	Ref.df	p
s(Target F0 range):LANG(EN)	2.5	3	0.2
s(Target F0 range):LANG(MD)	3.5	3.8	<0.001
s(Target F0 range, participant)	59.8	178	<0.001

Table A2: Summary of smooth terms of the GAM model for imitated intensity.

	Estimate	SE	p
(Intercept)	74.6	0.8	<0.001
LANG (MD-EN)	-1.4	1.2	0.23
	edf	Ref.df	p
s(Target F0 range):LANG(EN)	1	1	0.62
s(Target F0 range):LANG(MD)	2.2	2.6	0.003
s(Target F0 range, participant)	73.1	178	<0.001

Chapter 5

General discussion and conclusion

The broad goal of this dissertation is to provide insight into the mechanism of imitation. Through investigations of phonological mediation on phonetic imitation, we explored: (1) when and why F0 and duration show linear or non-linear imitation, (2) which levels of categories (linguistic level, psychophysical level, or a mix of both) can mediate imitation, and (3) how the distribution of imitations reflects the gradient nature of perception of phonological categories.

These questions were explored by examining the distribution of imitations of tonal continua: flat-falling tonal continua in Mandarin and checked-unchecked continua in Taiwanese Southern Min (TSM). The linearity and categoricalness of the imitation were analyzed by comparing the fit performance of two cognitive models: a linear regression model, which assumes speakers linearly track phonetic cues, and a categorical regression model, which assumes imitation reflects underlying categories. Specifically, the contribution of each model in accounting for variations in the distribution is interpreted as the degree of the linearity and categoricalness in the imitation.

Three studies were presented in this dissertation. Section 5.1 summarises the motivations and main findings of each study. Section 5.2 presents a general discussion of the results of all studies, and Section 5.3 concludes.

5.1 Summary

The first study, presented in Chapter 2, investigates the imitation of Mandarin flat-falling tonal continua by native speakers. The study manipulated F0 (range) and duration, which serve as primary and secondary cues to the tonal contrast, respectively. The main goal is to examine whether the imitation of these two cues is mediated by the tonal contrast, and whether their imitation exhibits a linear or categorical pattern. A categorization experiment and an imitation experiment were conducted using the same stimuli. The categorization experiment results showed that, as expected, duration was used as an additional cue (to the primary cue, F0 range) for discriminating the flat versus falling tonal contrast in Mandarin. Imitation of F0 and duration showed different patterns: the primary F0 range cue was imitated more categorically, while the less important duration cue was imitated more linearly. Additionally, although the continuously changing F0 ranges were imitated into two categories aligning with flat and falling tones, within the phonological falling category, the phonetic differences in the degree of F0 fall were reflected in the imitation. Furthermore, participants differed in the degree to which they categorically imitated the F0 and duration cues.

Study 2, presented in Chapter 3, explored syllable duration imitation through the checked-unchecked tonal contrast in TSM. This study was motivated by previous findings that duration imitation tends not to be mediated by phonological contrasts. However, those findings were based on cases where duration was a secondary cue to the phonological contrasts under investigation. TSM contains checked tones, which are typically short, and unchecked tones. The first experiment explored how important duration is in discriminating two checked-unchecked tonal contrasts in TSM, T3 vs.T33 and T5 vs. T55, and found that

duration is an important cue for the categorization of the T3-T33 contrast (but not for the T5-T55 contrast). The second experiment investigated the pattern of duration imitation in the T3-T33 tonal continuum, which was created by manipulating (only) duration. The results showed that the duration imitation was much more linear than categorical–the weight of the linear model is 92% whereas that of the categorical model is 8%, in contrast to the findings in Chapters 2 and 4 for F0 range imitation. This suggests that duration imitation is not mediated by the T3-T33 contrast, although it acts as a primary cue to it.

The third study, presented in Chapter 4, investigated the imitation of Mandarin flat-falling tonal continua by naïve English speakers and compared their imitation to that of native Mandarin speakers. The main goal is to explore whether speakers lacking lexical tone phonology imitate F0 range in a linear or categorical manner, in order to understand the type of contrast (linguistic versus psychophysical) that mediates phonetic imitation. The flat-falling tonal continuum was created by manipulating only F0 range. As in Study 1, a categorization experiment and an imitation experiment were conducted using the same continuum. Data from a group of native Mandarin speakers were collected, for comparison to native speaker performance. Categorization results showed that both English and Mandarin speakers exhibited a categorization effect in the identifying the tonal categories, replicating the findings of Chapter 2 and previous work. Particularly, English speakers' categorization was less categorical than that of native Mandarin speakers, and that they identified more stimuli as falling category than Mandarin speakers. In their imitation, results showed that both pre-existing categories and within-category phonetic details were involved in phonetic imitation, with the Mandarin group's imitation being more heavily influenced by the pre-existing categories compared to the English group – for Mandarin speakers' imitation, the weight of

the linear model is 41% whereas that of the categorical model is 59%; for English speakers' imitation, the weight of the linear model is 52% whereas that of the categorical model is 48% . Additionally, Mandarin speakers' imitations became shorter as the F0 range increased, whereas the durations of English speakers' imitations were not affected by the F0 range. Lastly, traces of hyper-articulation, characterized by larger F1 and intensity for larger F0 ranges, were observed in Mandarin speakers' imitations but not in those of English speakers. These discrepancies indicate distinct underlying categories mediating imitation in native and non-native groups: higher-level linguistic categories in the Mandarin group and lower-level psychophysical categories in the English group.

These results are discussed in the following sections.

5.2 General discussion

5.2.1 Implications for the phonological mediation effect on phonetic imitation

Results presented in this dissertation offer the following implications regarding the phonological mediation effect on phonetic imitation. First, phonological mediation in phonetic imitation is dimension-dependent – it is not specific to 'segmental' dimensions and does not apply to all dimensions uniformly. The influence of phonological categories leads to non-linear imitation of sounds that lie between the categories. Such phonological mediation of imitation has been previously observed in dimensions including VOT (in stop imitation: Flege & Eefting, 1988; Nielsen, 2011; Mitterer & Ernrstus. 2008) and formants (in vowel imitation: Chistovich et al., 1966; Kent, 1974; Repp, 1984; Kim & Clayards, 2019). However,

for dimensions primarily used for marking suprasegmental contrasts – F0, duration and intensity – phonological mediation has been less examined, and results have been mixed in previous work. Study 1 confirmed that, like VOT and formants for segmental categories, the imitation of F0 range for productions between tonal categories (i.e., the flat and falling tonal categories in Mandarin) is mediated by category structure. However, Study 2 found that in a case where duration is an important cue to a tonal contrast (i.e., the T3 and T33 tonal categories in TSM), its imitation was *not* mediated by category structure. The differing imitation patterns of F0 and duration show that phonological effects on imitation are dimension -dependent. Importantly, these findings challenge the division between 'segmental' dimensions and 'supra-segmental' dimensions in whether imitation is influenced by phonological contrasts versus linear. These findings underscore the need for a better understanding of how phonological effects on imitation differ by dimension (e.g., duration versus F0), which is crucial for a more nuanced comprehension of the mechanisms involved in phonetic imitation and its role in language processing.

Second, this dissertation demonstrates that phonetic imitation is not solely mediated by stronger, higher-level phonological categories, but also by weaker, lower-level psychophysical categories. Study 3 showed that English speakers who lacked experience with lexical tone phonology did not imitate the flat-falling F0 continuum linearly. Instead, their imitations showed influence of a flat category and a falling category, similarly to native Mandarin speakers. This result clearly indicates that phonetic imitation is mediated by categories beyond just phonological ones, which cannot be at play for the English speakers when perceiving a tonal contrast. It has been hypothesized that when categorizing the flat-falling tonal contrasts, naïve English speakers are using psychophysical categories which generally distinguish rising

or falling from flat F0 contours (Wang, 1976). Study 3 provided evidence supporting this view. Mandarin speakers showed imitations that contained variation in duration, F1 and intensity – consistent with hyper-articulation induced by the phonology of the Mandarin flat vs. falling tonal contrast. In contrast, English speakers' imitation was limited to the F0 dimension. This finding suggests that the categories involved in English speakers' imitation were lower-level, resulting in imitated productions which did not vary in other dimensions (i.e., secondary cues to the tonal contrast). To our knowledge, this is the first study showing that phonetic imitation is mediated by both linguistic and psychophysical contrasts. These findings also shed light on the influence of both linguistic and psychophysical contrasts in speech perception and production.

5.2.2 Imitation as an experimental paradigm

Despite observing a mostly categorical imitation pattern for F0 range, we found that within-category phonetic details were also maintained in imitation, especially for the falling category. Both native Mandarin speakers (in Study 1 and Study 2) and English speakers (in Study 3) tracked the degree of F0 fall when producing imitations influenced by the falling tone category, while also exhibiting a flat category. Similar within-category effects in imitation have also been observed for stop VOT, where imitation of shortened, canonical and lengthened VOT have all been observed (Nielsen, 2011; Schertz and Johnson, 2022; Schertz et al., 2023). Our findings are consistent with a growing body of studies that highlight the gradient nature of speech perception, alongside categorical effects (see McMurray, 2022 for a review).

Whether a more categorical/phonological or more gradient/phonetic pattern is observed depends on the experimental paradigm and task demands (Strange, 2011). While

categorization experiments, such as the 2AFC identification paradigm, can indicate the degree of categoricalness and the categorization boundary position, they provide limited information about the actual structure of each category involved in the contrast. In contrast, as shown in these three studies, phonetic imitation of the same continuum used for the categorization task offers more insights into category structure, similar to how the goodness rating task does in perceptual experiments (Kuhl, 1991; Miller and Volaitis, 1989; Strange et al., 2004). The studies in this dissertation highlight that phonetic imitation is a paradigm that taps into both the phonetic (continuous) nature and phonological (categorical) nature of speech processing, both of which are observable in speakers' imitative behavior.

Furthermore, the strength of the phonological mediation effect can be flexible, depending on factors such as sentence context. The mediation effect seems to be stronger in Zhang et al. (2023) than in the present study, as reflected by a more categorical F0 imitation distribution compared to the present study. This could be due to Zhang et al. (2023) using a carrier phrase in the imitation task, whereas the present study only included the target syllable. Previous research (e.g., Francis et al., 2003) has shown that context can enhance the categorization effect in the perception of Mandarin tones. The different levels of discreteness observed between Zhang et al. (2023) and the current study indicate that this context effect may extend from perception to production in an imitation task.

5.2.3 Implications for the examined tonal contrasts

This dissertation examined imitation of the Mandarin flat-falling tone contrast and the high-register checked-unchecked tone contrast in TSM, focusing on F0 and duration, whose role

as primary versus secondary cue differs between the two cases. The results have implications for our understanding of these two tonal contrasts.

Regarding Mandarin tones, previous studies have shown mixed results on whether duration influences tonal categorization (Wang and Peng, 2012; Feng and Peng, 2018; Zhu et al., 2016; Wang et al., 2017; Yang, 1989; Blicher et al., 1990). However, Study 1 clearly showed that duration played a role in the flat-falling contrast. In the categorization task, shorter duration increased the likelihood of a falling tone response. In the imitation task, a shorter duration increased the probability of the imitation belonging to the falling category. The discrepancy between the current findings and some previous studies (e.g., Wang and Peng, 2012; Wang et al., 2017) that failed to observe the duration effect on the flat-falling contrast could be due to differing values of duration and F0 range used in these experiments. Since the effect of duration is secondary to the flat-falling contrast, it may be sensitive to the specific duration and F0 range values used. The current study observed a clear effect of duration on the tonal contrast in both perception and imitation tasks, possibly due to the relatively larger range of F0 falls and duration values used.

TSM is undergoing sound change in its tonal system, and the results from Study 3 may reflect this. Previous studies have shown that the two checked tones are merging and the final codas are being lost (Ang, 2013; Pan, 2017), suggesting a decreasing phonological awareness of the checked-unchecked quality. Additionally, there seems to be an increasing awareness of durational quantity. As confirmed in Study 3, the role of duration in distinguishing checked vs. unchecked tones is pronounced for the mid register.

5.2.4 Future directions

The present dissertation explores the impact of phonology and phonetics on imitation of tones cued by F0 and vowel duration. This research contributes to our understanding of the mechanism of phonetic imitation and provides insight into theories of speech processing by showing both categorical and continuous aspects of speech perception, as well as the involvement of both linguistic and psychophysical categories in speech perception. The findings of the three studies also suggest the following directions for future research.

This dissertation primarily focused on how categorical imitation looked at the group level (e.g., Mandarin vs. English speakers), however, we found significant variation among participants in how linear their imitations were versus influenced by tonal categories. Interestingly, there seemed to be participants whose imitation of F0 was not affected by the Mandarin tone contrast (as observed in Study 1). Future studies should investigate the factors affecting individual differences in linearity versus phonological mediation of imitation.

In addition to individual differences, we found that native speakers' F0 imitation in Study 1 seemed be more categorical than those in Study 2. This could stem from differences in the stimuli used in the two studies. First, Study 1 employed a semantically neutral carrier (meaning "the sound of") in the imitation task, while Study 2 used isolated syllables. The lack of context may strengthen the auditory processing mode, resulting in more gradient imitation behavior. Second, Study 2 used a longer duration for the target syllable than Study 1. Longer duration may provide time for more cognitive resources to be used for the perception task, and allow participants the opportunity for finer control of production (in the imitation task). Third, there were 9 steps for F0 range in Study 2 whereas there were 7 in Study 1. As the maximum F0 fall was similar across the two studies, Study 2 had a smaller step size. This

might have given us more resolution to observe gradient behaviour. Future studies could examine the extent to which each of these factors influences imitation.

5.3 Conclusion

Phonological mediation has been found in imitation of segmental dimensions, while imitation of suprasegmental dimensions has been less studied and results are mixed. This dissertation examined phonological effects on the imitation of F0 and duration in the context of two lexical tone contrasts.

First, the results highlight the dimension-dependent nature of phonological effects on the degree to which imitation is linear. While phonological mediation was evident for F0 range, it was not observed for duration, even in a case where duration is the primary cue (Study 3), suggesting that phonological mediation is not uniform across all dimensions. This finding also challenges the division between 'segmental' dimensions and 'supra-segmental' dimensions in terms of their susceptibility to phonological mediation in imitation.

Secondly, this research reveals that phonetic imitation is not solely mediated by phonological categories, but also by psychophysical categories. English speakers lacking experience with lexical tone phonology imitated F0 range variation in a categorical manner, suggesting the involvement of lower-level psychophysical categories in phonetic imitation.

Furthermore, within-category phonetic details influenced imitation, a manifestation of the gradient nature of speech perception. The role of both categories and within-category phonetic variation highlights how phonetic imitation is a paradigm where both the phonetic (continuous) nature and phonological (categorical) nature of speech processing play a role.

Overall, the findings from this dissertation provide insight into the mechanism of phonetic imitation and contribute to theories of speech processing by showing both categorical and continuous aspects of speech perception, as well as the involvement of both linguistic and psychophysical categories in speech perception.

References

Adank, Patti, Peter Hagoort, and Harold Bekkering. 2010. "Imitation Improves Language Comprehension." Psychological Science 21(12):1903–9. doi: 10.1177/0956797610389192.

Adank, Patti, Andrew J. Stewart, Louise Connell, and Jeffrey Wood. 2013. "Accent Imitation Positively Affects Language Attitudes." Frontiers in Psychology 4. doi: 10.3389/fpsyg.2013.00280.

Ang, Uijin. 2013. "The Distribution and Regionalization of Varieties in Taiwan." Language and Linguistics 14(2):315.

Babel, Molly Elizabeth. 2009. Phonetic and Social Selectivity in Speech Accommodation. University of California, Berkeley.

Beckman, Mary E. 1996. "The Parsing of Prosody." Language and Cognitive Processes 11(1–2):17–68.

Bien, N., A. Roebroeck, R. Goebel, and A. T. Sack. 2009. "The Brain's Intention to Imitate: The Neurobiology of Intentional versus Automatic Imitation." Cerebral Cortex 19(10):2338–51. doi: 10.1093/cercor/bhn251.

Blicher, Deborah L., Randy L. Diehl, and Leslie B. Cohen. 1990. "Effects of Syllable Duration on the Perception of the Mandarin Tone 2/Tone 3 Distinction: Evidence of Auditory Enhancement." Journal of Phonetics 18(1):37–49. doi: 10.1016/S0095-4470(19)30357-2.

Brophy, Timothy S. 2001. "Developing Improvisation in General Music Classes." Music Educators Journal 88(1):34–53.

Chartrand, Tanya L., and John A. Bargh. 1999. "The Chameleon Effect: The Perception–Behavior Link and Social Interaction." Journal of Personality and Social Psychology 76(6):893.

Chistovich, L., G. Fant, A. de Serpa-Leitao, and P. Tjernlund. 1966. "Mimicking of Synthetic Vowels." Quarterly Progress and Status Report, Speech Transmission Lab, Royal Institute of Technology, Stockholm 1(2):1–18.

Coupland, Justine, Nikolas Coupland, and Howard Giles. 1991. "Accommodation Theory. Communication, Context and Consequences." Contexts of Accommodation 1–68.

Custance, Deborah M., Kim A. Bard, and Andrew Whiten. 1995. "Can Young Chimpanzees (Pan Troglodytes) Imitate Arbitrary Actions? Hayes & Hayes (1952) Revisited." Behaviour 132(11–12):837–59.

Cutler, Anne, Delphine Dahan, and Wilma van Donselaar. 1997. "Prosody in the Comprehension of Spoken Language: A Literature Review." Language and Speech 40(2):141–201. doi: 10.1177/002383099704000203.

De Meo, Anna, Marilisa Vitale, Massimo Pettorino, Francesco Cutugno, and Antonio Origlia. 2013. "Imitation/Self-Imitation in Computer-Assisted Prosody Training for Chinese Learners of L2 Italian." Pronunciation in Second Language Learning and Teaching Proceedings 4(1).

Delvaux, Véronique, and Alain Soquet. 2007. "The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation." Phonetica 64(2–3):145–73. doi: 10.1159/000107914.

Dilley, Laura C. 2010. "Pitch Range Variation in English Tonal Contrasts: Continuous or Categorical?" Phonetica 67(1–2):63–81. doi: 10.1159/000319379.

Dimberg, Ulf, Monika Thunberg, and Kurt Elmehed. 2000. "Unconscious Facial Reactions to Emotional Facial Expressions." Psychological Science 11(1):86–89. doi: 10.1111/1467-9280.00221.

D'Imperio, Mariapaola, and James Sneed. 2015. "Phonetic Detail and the Role of Exposure in Dialect Imitation." in 18 th International Congress of Phonetic Sciences.

Elert, C. C. 1964. "Phonologic Studies of Swedish Quantity." Uppsala: Almqvist & Wiksell.

Feng, Yan, and Gang Peng. 2018. "The Effect of Duration on Categorical Perception of Mandarin Tone and Voice Onset Time." Pp. 164–68 in TAL2018, Sixth International Symposium on Tonal Aspects of Languages. ISCA.

Fitch, W. Tecumseh. 2000. "The Evolution of Speech: A Comparative Review." Trends in Cognitive Sciences 4(7):258–67.

Flege, James Emil, and Wieke Eefting. 1988. "Imitation of a VOT Continuum by Native Speakers of English and Spanish: Evidence for Phonetic Category Formation." The Journal of the Acoustical Society of America 83(2):729–40. doi: 10.1121/1.396115.

Fowler, Carol A., Julie M. Brown, Laura Sabadini, and Jeffrey Weihing. 2003. "Rapid Access to Speech Gestures in Perception: Evidence from Choice and Simple Response Time Tasks." Journal of Memory and Language 49(3):396–413. doi: 10.1016/S0749-596X(03)00072-X.

Francis, A. L., Ciocca, V., & Chit Ng, B. K. (2003). On the (non)categorical perception of lexical tones. Perception & Psychophysics, 65(7), 1029–1044. https://doi.org/10.3758/BF03194832

Giles, Howard. 1973. "Accent Mobility: A Model and Some Data." Anthropological Linguistics 87–105.

Goldinger, Stephen D. 1998. "Echoes of Echoes? An Episodic Theory of Lexical Access." Psychological Review 105(2):251.

Goldsmith, John A. 1990. Autosegmental and Metrical Phonology. Vol. 1. Basil Blackwell.

Heyes, C. 2001. "Causes and Consequences of Imitation." Trends in Cognitive Sciences 5(6):253–61. doi: 10.1016/S1364-6613(00)01661-2.

Hillenbrand, James, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler. 1995. "Acoustic Characteristics of American English Vowels." The Journal of the Acoustical Society of America 97(5):3099–3111.

Holley, Freda M., and Janet K. King. 1971. "Imitation and Correction in Foreign Language Learning." The Modern Language Journal 55(8):494–98.

Iacoboni, Marco, Roger P. Woods, Marcel Brass, Harold Bekkering, John C. Mazziotta, and Giacomo Rizzolatti. 1999. "Cortical Mechanisms of Human Imitation." Science 286(5449):2526–28. doi: 10.1126/science.286.5449.2526.

Jensen, Eva Dam, and Thora Vinther. 2003. "Exact Repetition as Input Enhancement in Second Language Acquisition: Language Learning." Language Learning 53(3):373–428. doi: 10.1111/1467-9922.00230.

Kapnoula, Efthymia C., and Bob McMurray. 2021. "Idiosyncratic Use of Bottom-up and Top-down Information Leads to Differences in Speech Perception Flexibility: Converging Evidence from ERPs and Eye-Tracking." Brain and Language 223:105031. doi: 10.1016/j.bandl.2021.105031.

Kent, R. D. 1974. "Auditory-Motor Formant Tracking: A Study of Speech Imitation." Journal of Speech and Hearing Research 17(2):203–22. doi: 10.1044/jshr.1702.203.

Kent, Raymond D. 1973. "The Imitation of Synthetic Vowels and Some Implications for Speech Memory." Phonetica 28(1):1–25.

Kim, Donghyun, and Meghan Clayards. 2019. "Individual Differences in the Link between Perception and Production and the Mechanisms of Phonetic Imitation." Language, Cognition and Neuroscience 34(6):769–86. doi: 10.1080/23273798.2019.1582787.

Kovaříková, Anna. 2023. "Subconscious Imitation of Phonetic Features of Perceived Speech and Its Influence on Phonological Contrast." Univerzity Palackého.

Kuhl, Patricia K. 1991. "Human Adults and Human Infants Show a 'Perceptual Magnet Effect' for the Prototypes of Speech Categories, Monkeys Do Not." Perception & Psychophysics 50(2):93–107. doi: 10.3758/BF03212211.

Kuhl, Patricia K., and Andrew N. Meltzoff. 1982. "The Bimodal Perception of Speech In Infancy." Science 218(4577):1138–41. doi: 10.1126/science.7146899.

Kuhl, Patricia K., and Andrew N. Meltzoff. 1996. "Infant Vocalizations in Response to Speech: Vocal Imitation and Developmental Change." The Journal of the Acoustical Society of America 100(4):2425–38. doi: 10.1121/1.417951.

Kwon, Harim. 2019. "The Role of Native Phonology in Spontaneous Imitation: Evidence from Seoul Korean." Laboratory Phonology: Journal of the Association for Laboratory Phonology 10(1):10. doi: 10.5334/labphon.83.

Lehiste, Ilse, and Gordon E. Peterson. 1959. "Vowel Amplitude and Phonemic Stress in American English." The Journal of the Acoustical Society of America 31(4):428–35.

Lewandowski, Natalie, and Matthias Jilka. 2019. "Phonetic Convergence, Language Talent, Personality and Attention." Frontiers in Communication 4.

Liberman, Alvin M., Katherine Safford Harris, Howard S. Hoffman, and Belver C. Griffith. 1957. "The Discrimination of Speech Sounds within and across Phoneme Boundaries." Journal of Experimental Psychology 54(5):358.

Lin, Yuhan, Yao Yao, and Jin Luo. 2021. "Phonetic Accommodation of Tone: Reversing a Tone Merger-in-Progress via Imitation." Journal of Phonetics 87:101060. doi: 10.1016/j.wocn.2021.101060.

MacLeod, Bethany. 2015. "A Critical Evaluation of Two Approaches to Defining Perceptual Salience." Ampersand 2:83–92. doi: 10.1016/j.amper.2015.07.001.

MacLeod, Bethany, and Sabrina M. Di Lonardo Burr. 2022. "Phonetic Imitation of the Acoustic Realization of Stress in Spanish: Production and Perception." Journal of Phonetics 92:101139. doi: 10.1016/j.wocn.2022.101139.

Massaro, Dominic W., and Michael M. Cohen. 1983. "Categorical or Continuous Speech Perception: A New Test." Speech Communication 2(1):15–35. doi: 10.1016/0167-6393(83)90061-4.

McMurray, Bob. 2022. "The Myth of Categorical Perception." The Journal of the Acoustical Society of America 152(6):3819–42. doi: 10.1121/10.0016614.

McMurray, Bob, Richard N. Aslin, Michael K. Tanenhaus, Michael J. Spivey, and Dana Subik. 2008. "Gradient Sensitivity to Within-Category Variation in Words and Syllables." Journal

of Experimental Psychology: Human Perception and Performance 34(6):1609–31. doi: https://doi.org/10.1037/a0011747.

McMurray, Bob, Michael K. Tanenhaus, and Richard N. Aslin. 2002. "Gradient Effects of Within-Category Phonetic Variation on Lexical Access." Cognition 86(2):B33–42. doi: 10.1016/S0010-0277(02)00157-9.

Meltzoff, Andrew N. 2011. "Social Cognition and the Origins of Imitation, Empathy, and Theory of Mind." 49–75.

Meltzoff, Andrew N., and M. Keith Moore. 1977. "Imitation of Facial and Manual Gestures by Human Neonates." Science 198(4312):75–78. doi: 10.1126/science.198.4312.75.

Meltzoff, Andrew N., and M. Keith Moore. 1997. "Explaining Facial Imitation: A Theoretical Model." Infant and Child Development 6(3–4):179–92.

Miller, Joanne L., and Lydia E. Volaitis. 1989. "Effect of Speaking Rate on the Perceptual Structure of a Phonetic Category." Perception & Psychophysics 46(6):505–12. doi: 10.3758/BF03208147.

Mitterer, Holger, and Mirjam Ernestus. 2008. "The Link between Speech Perception and Production Is Phonological and Abstract: Evidence from the Shadowing Task." Cognition 109(1):168–73. doi: 10.1016/j.cognition.2008.08.002.

Niedzielski, Nancy, and Howard Giles. 1996. "Linguistic accommodation." Pp. 332–42 in Linguistic accommodation. De Gruyter Mouton.

Nielsen, Kuniko. 2011. "Specificity and Abstractness of VOT Imitation." Journal of Phonetics 39(2):132–42. doi: 10.1016/j.wocn.2010.12.007.

Nielsen, Kuniko. 2014. "Phonetic Imitation by Young Children and Its Developmental Changes." Journal of Speech, Language, and Hearing Research 57(6):2065–75. doi: 10.1044/2014_JSLHR-S-13-0093.

Nielsen, Kuniko Y., and Rebecca Scarborough. 2015. "Perceptual Asymmetry between Greater and Lesser Vowel Nasality and VOT." in ICPhS.

Pan, Ho-hsien. 2017. "Glottalization of Taiwan Min Checked Tones." Journal of the International Phonetic Association 47(1):37–63. doi: 10.1017/S0025100316000281.

Paquette-Smith, Melissa, Jessamyn Schertz, and Elizabeth K. Johnson. 2022. "Comparing Phonetic Convergence in Children and Adults." Language and Speech 65(1):240–60. doi: 10.1177/00238309211013864.

Pardo, Jennifer S. 2006. "On Phonetic Convergence during Conversational Interaction." The Journal of the Acoustical Society of America 119(4):2382–93.

Pardo, Jennifer S. 2010. "Expressing Oneself in Conversational Interaction." 14.

Pardo, Jennifer S., Adelya Urmanche, Sherilyn Wilman, and Jaclyn Wiener. 2017. "Phonetic Convergence across Multiple Measures and Model Talkers." Attention, Perception, & Psychophysics 79(2):637–59.

Pfordresher, Peter Q., James T. Mantell, and Tim A. Pruitt. 2022. "Effects of Intention in the Imitation of Sung and Spoken Pitch." Psychological Research 86(3):792–807. doi: 10.1007/s00426-021-01527-0.

Pickering, Martin J., and Simon Garrod. 2004. "Toward a Mechanistic Psychology of Dialogue." Behavioral and Brain Sciences 27(02). doi: 10.1017/S0140525X04000056.

Pickering, Martin J., and Simon Garrod. 2013. "An Integrated Theory of Language Production and Comprehension." Behavioral and Brain Sciences 36(4):329–47. doi: 10.1017/S0140525X12001495.

Pierrehumbert, Janet B., and Shirley A. Steele. 1989. "Categories of Tonal Alignment in English." Phonetica 46(4):181–96.

Repp, Bruno H. 1984. "Categorical Perception: Issues, Methods, Findings." Pp. 243–335 in Speech and language. Vol. 10. Elsevier.

Rizzolatti, Giacomo, Luciano Fadiga, Vittorio Gallese, and Leonardo Fogassi. 1996. "Premotor Cortex and the Recognition of Motor Actions." Cognitive Brain Research 3(2):131–41.

Rizzolatti, Giacomo, Leonardo Fogassi, and Vittorio Gallese. 2001. "Neurophysiological Mechanisms Underlying the Understanding and Imitation of Action." Nature Reviews Neuroscience 2(9):661–70. doi: 10.1038/35090060.

Sancier, Michele L., and Carol A. Fowler. 1997. "Gestural Drift in a Bilingual Speaker of Brazilian Portuguese and English." Journal of Phonetics 25(4):421–36.

Sato, Marc, Krystyna Grabski, Maëva Garnier, Lionel Granjon, Jean-Luc Schwartz, and Noël Nguyen. 2013. "Converging toward a Common Speech Code: Imitative and Perceptuo-Motor Recalibration Processes in Speech Production." Frontiers in Psychology 4. doi: 10.3389/fpsyg.2013.00422.

Schertz, Jessamyn, and Elizabeth K. Johnson. 2022. "Voice Onset Time Imitation in Teens Versus Adults." Journal of Speech, Language, and Hearing Research 65(5):1839–50. doi: 10.1044/2022_JSLHR-21-00460.

Schertz, Jessamyn, and Melissa Paquette-Smith. 2023. "Convergence to Shortened and Lengthened Voice Onset Time in an Imitation Task." JASA Express Letters 3(2):025201. doi: 10.1121/10.0017066.

Strange, Winifred, Ocke-Schwen Bohn, Sonja A. Trent, and Kanae Nishi. 2004. "Acoustic and Perceptual Similarity of North German and American English Vowels." The Journal of the Acoustical Society of America 115(4):1791–1807. doi: 10.1121/1.1687832.

Toscano, Joseph C., Bob McMurray, Joel Dennhardt, and Steven J. Luck. 2010. "Continuous Perception and Graded Categorization: Electrophysiological Evidence for a Linear Relationship between the Acoustic Signal and Perceptual Encoding of Speech." Psychological Science 21(10):1532–40.

Vallabha, G. K., & Tuller, B. (2004). Perceptuomotor bias in the imitation of steady-state vowels. J. Acoust. Soc. Am., 116(2).

Wang, Dazuo, and Gang Peng. 2012. "Effects of Pitch Range and Duration on Tone Categorical Perception." Retrieved June 12, 2021 (https://www.isca-speech.org/archive/tal_2012/tl12_O2-03.html).

Wang, William Shi Yuan. 1976. "Language Change." Annals of the New York Academy of Sciences.

Wang, Yuxia, Xiaohu Yang, and Chang Liu. 2017. "Categorical Perception of Mandarin Chinese Tones 1–2 and Tones 1–4: Effects of Aging and Signal Duration." Journal of Speech, Language, and Hearing Research 60(12):3667–77. doi: 10.1044/2017_JSLHR-H-17-0061.

Watkins, K. E., A. P. Strafella, and T. Paus. 2003. "Seeing and Hearing Speech Excites the Motor System Involved in Speech Production." Neuropsychologia 41(8):989–94. doi: 10.1016/S0028-3932(02)00316-0.

Whalen, D. H., Arthur S. Abramson, Leigh Lisker, and Maria Mody. 1993. "F0 Gives Voicing Information Even with Unambiguous Voice Onset Times." The Journal of the Acoustical Society of America 93(4):2152–59. doi: 10.1121/1.406678.

Wilson, Stephen M., Ayşe Pinar Saygin, Martin I. Sereno, and Marco Iacoboni. 2004. "Listening to Speech Activates Motor Areas Involved in Speech Production." Nature Neuroscience 7(7):701–2. doi: 10.1038/nn1263.

Wolpert, Daniel M., Zoubin Ghahramani, and J. Randall Flanagan. 2001. "Perspectives and Problems in Motor Learning." Trends in Cognitive Sciences 5(11):487–94. doi: 10.1016/S1364-6613(00)01773-3.

Yang, Y. (1989). The vowels and the perception of Chinese tones. ACTA Psychologica Sinica (1): 29-33. (In Chinese)

Yu, Alan C. L., Carissa Abrego-Collier, and Morgan Sonderegger. 2013. "Phonetic Imitation from an Individual-Difference Perspective: Subjective Attitude, Personality and 'Autistic' Traits" edited by J. Snyder. PLoS ONE 8(9):e74746. doi: 10.1371/journal.pone.0074746.

Zellou, Georgia, Delphine Dahan, and David Embick. 2017. "Imitation of Coarticulatory Vowel Nasality across Words and Time." Language, Cognition and Neuroscience 32(6):776–91. doi: 10.1080/23273798.2016.1275710.

Zellou, Georgia, Rebecca Scarborough, and Kuniko Nielsen. 2016. "Phonetic Imitation of Coarticulatory Vowel Nasalization." The Journal of the Acoustical Society of America 140(5):3560–75. doi: 10.1121/1.4966232.

Zetterholm, Elisabeth. 2002. "A Case Study of Successful Voice Imitation." Logopedics Phoniatrics Vocology 27(2):80–83.

Zetterholm, Elisabeth. 2007. "Detection of Speaker Characteristics Using Voice Imitation." Pp. 243–57 in Speaker Classification II. Vol. 4441, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg.

Zhu, W., Wei Y., Wu L. & Wang J. (2016). The Effect of Pitch Range and Tone Duration on Chinese Tone Perception by L2 Learners.Chinese Language Learning (2): 83-92 (In Chinese).