Cognitive Factors in Perception and Imitation of Thai Tones by Mandarin versus Vietnamese Speakers

Juqiang Chen

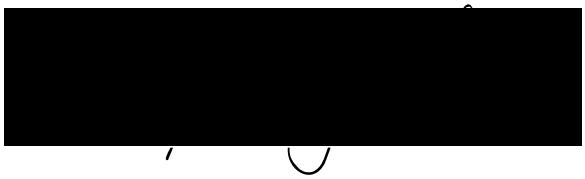
Thesis submitted for the degree of Doctor of Philosophy

Western Sydney University

2020

I hereby declare that this submission is my own work and, to the best of my knowledge, it contains no material previously published or written by any other person, nor material which has been accepted for the award of any other degree or diploma at Western Sydney University, or any other educational institution, except where due acknowledgement is made in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception is acknowledged.



Acknowledgements

First, I would like to thank my primary supervisor Professor Catherine Best for her guidance, support, and encouragement from my earlier application stage and throughout the whole candidature. Many a time I stumbled and felt at sea in the face of unexpected results. Her profound knowledge, in-depth thinking and unfailing support offered a beacon of hope.

Next, to my panel member and supervisor, Dr. Mark Antoniou, his expertise in Praat and E-prime programming and experiment design saved me a lot of time and made my learning curve of these less steep. I am also grateful to his practical advice on many aspects of career planning and life. To Dr. Benjawan Kasisopa, I am indebted to her help in preparing the Thai stimuli and experiment design. To Professor Denis Burnham, I am thankful for allowing me to use some data from his Thai tone corpus. To Dr. Nan Xu and Dr. Ivan Yuen, thank you for letting me use the acoustic lab for testing at Macquarie University. To Dr. Xuliang He, thank you for introducing me to phonetic research. To Professor Hua Chen, for all the valuable academic and life advice and encouragement. To the MARCS staff members, Darlene Williams, Krista Arief, Karen McConachie and Jessica Simpson, thank you for your support at every stage of my candidature and all the academic travels. To the technique team, Ben Binyamin, Dr Leidy Castro-Meneses, Johnson Chen, Dr Chris Wang, thanks for your help with all the equipment and programming issues.

To all the students and colleagues at MARCS, it is my great honour and pleasure to meet and work with you. Special thanks to Tina Whyte-Ball, Yassine Frej, Yanping Li for their comments at the lab meetings and for friendship and lots of laughs.

To all my family and friends, past and present, home and abroad, thank you for being there when I am in need.

Table of Contents

List of tables

List of figures

Abstract

The thesis investigates how native language phonological and phonetic factors affect non-native lexical tone perception and imitation, and how cognitive factors, such as memory load and stimulus variability (talker and vowel context variability), bias listeners to a phonological versus phonetic mode of perception/imitation. Two perceptual experiments and one imitation experiment were conducted with Thai tones as the stimuli and with Mandarin and Vietnamese listeners, who had no experience with Thai (i.e., naïve listeners/imitators). The results of the perceptual experiments (Chapters 5 and 6) showed phonological effects as reflected in assimilation types (Categorised vs. UnCategorised assimilation) and phonetic effects indicated by percent choice and goodness ratings in tone assimilation, largely in line with predictions based on the Perceptual Assimilation Model (PAM: Best, 1995). In addition, phonological assimilation types and phonological overlap of the contrasts affected their discrimination in line with predictions based on PAM. Phonetic difference scores calculated with percent choice and goodness ratings distinguished between non-native contrasts of the same phonological assimilation type and overlap type, clarifying the role of residual phonetic sensitivity in non-native discrimination. Discrimination accuracy was reduced by high talker and vowel variability, which bias listeners to use a phonological mode of perception according to principles of the Automatic Selective Perception model (ASP: Strange, 2011), but was unaffected by memory load manipulations. On the other hand, assimilated responses were influenced by memory load and showed a more phonological-based pattern, i.e., higher percent choice and/or goodness ratings, under high than low memory load. Assimilation responses were unaffected by talker and vowel variability because the assimilation task itself requires use of the phonological mode.

Differences in deviation scores in the imitation experiment (Chapter 7) between the two language groups can be accounted for with reference to their respective perceptual assimilation patterns. Although native phonological influences constrain non-native tone imitation, imitators retained phonetic sensitivity to the specific details of the target stimuli in line with their percent choice/goodness ratings in tone assimilations. Results support PAM's principles that native phonological and phonetic effects on non-native speech perception extend to non-native tone imitation. Both Mandarin and Vietnamese imitators produced more accurate imitation under low memory load and in constant talker blocks where phonetic mode of perception/imitation was activated and phonetic details in the stimuli were available in working memory and were attended to, compared with when the imitators were under high memory load and in variable talker blocks.

The thesis research has revealed the influence of cognitive factors on native language influences in perception and imitation of non-native lexical tones, which contribute differently to different tasks. The findings carry implications for current non-native speech perception theories. The fact that non-native tone imitation deviations can be traced back to native phonological and phonetic influences on perception supports and provides new insights about perception-production links in processing non-native tones. The findings uphold the extrapolation of PAM and ASP principles to non-native tone perception and imitation, indicating that both native language phonological and phonetic influences and their modulation by cognitive factors hold implications for non-native speech perception/learning theories, as well as for second language instruction.

Chapter 1. Overview

This thesis investigates how native language phonological and phonetic factors affect non-native perception and imitation of Thai tones by Mandarin and Vietnamese listeners, who had no experience with Thai, i.e., naïve listeners/imitators, and how cognitive factors bias listeners toward phonetic versus phonological modes of perception. Hypotheses and predictions on non-native lexical tone perception and imitation performances were framed in light of cross-language speech perception/second language speech learning theories and were tested in three experiments.

Two introduction chapters (Chapter 2-3) were written to provide a solid theoretical base for the three experiments to follow (Chapter 5-7). Chapter 2 starts by reviewing major theoretical approaches to speech perception and their assumptions with a focus on how they address the lack of invariants issue in speech perception, i.e., the problem caused by different variations, such as, talker variability and phonetic contexts. Then, four non-native speech perception/learning theories that account for native language influence on non-native perception and/or production are reviewed: the Perceptual Assimilation Model (PAM: Best, 1995), the Speech Learning Model (SLM: Flege, 1995), the Native Language Magnet Model (NLM: Kuhl & Iverson, 1995; Kuhl, 1994) and the Second Language Linguistic Perception model (L2LP: Escudero, 2005). In addition, the Automatic Selective Perception model (ASP, Strange, 2011) is introduced to account for varying performance in speech perception in terms of two different modes of perception, i.e., phonological versus phonetic mode. Studies on how memory load and talker and/or vowel context variability shift listeners between two modes of perception are reviewed. These theories are compared in terms of their scopes and assumptions. Consequently, PAM is selected as the major theoretical framework for experimental chapters in this thesis as its principles consider both native

phonological and phonetic factors in perceptual assimilation and discrimination of non-native tones, and its focus on perception of articulatory information in speech allows extension to predict imitation of non-native tones. ASP is used in tandem with PAM to account for cognitive factors, which cause listeners/imitators to shift between a phonological and a phonetic mode of perception. Other theories, such as SLM and NLM, are discussed throughout the thesis wherever relevant.

Chapter 3 provides a review of research on non-native lexical tone perception and production. The chapter starts with an introduction of tonogenesis: the birth of tones in the three languages, that were used in the three experiments, i.e., Thai, Mandarin, and Vietnamese. Phonological features and phonetic characteristics of lexical tones are introduced. This is followed by a review of native and non-native tone perception and production studies to lay the foundation for the three experimental chapters: Chapters 5, 6 and 7.

Chapter 4 outlines the research framework and hypotheses for the subsequent experimental chapters that examine native phonological and phonetic influences on perception and imitation of Thai tones by Mandarin and Vietnamese speakers, and how those effects are modulated by cognitive factors, i.e., memory load, talker variability and vowel variability.

The three experimental chapters are the core of the thesis: Chapter 5 and 6 are perception experiments and Chapter 7 is an imitation experiment. These chapters are presented in the form of journal articles, as they were submitted to academic journals, with the exception that they have been modified to ensure consistency in formatting with the rest of the thesis.

In Chapter 5, we report a perceptual assimilation experiment that examined how native language phonological and phonetic factors affect the assimilation of Thai tones by Thai-naïve Mandarin, Northern Vietnamese, and Southern Vietnamese native listeners. This experiment tested PAM predictions on assimilation that consider both native phonological (assimilation types) and

phonetic factors (percent choice and goodness ratings). In addition, this experiment established the perceptual assimilation patterns of Thai tones by Mandarin and Vietnamese listeners upon which PAM predictions about discrimination of non-native tone contrasts were based. These predictions were used to select contrasts for testing in the second experiment (Chapter 6).

The experiments in Chapter 6 investigated how discrimination and assimilation of Thai tones by Mandarin and Vietnamese listeners are affected by native language phonological and phonetic factors. At the same time, cognitive factors, i.e., memory load, talker variability and vowel variability, were systematically manipulated in ways that were expected to shift listeners between a phonological and a phonetic mode of perception and consequently affect discrimination and assimilation performance.

In Chapter 7, imitation of Thai tones by Mandarin and Vietnamese listeners was tested to explore the relationship between perception and production. Native language phonological and phonetic influences on imitation were examined with reference to the perceptual assimilation patterns, i.e., assimilation type, percent choice and/or goodness ratings in Chapter 6. The same cognitive factors as in Chapter 6 were manipulated in Chapter 7, as they were expected to bias listeners to a phonological and phonetic mode of perception and thereby affect imitation performance.

The thesis concludes with a general discussion in Chapter 8 in which the experimental findings of Chapters 5-7 are summarised with regard to native phonological and phonetic factors on perception and imitation, as modulated by cognitive factors that lead to a switch between phonological and phonetic modes of perception. This chapter also situates the findings of the three experimental chapters in the context of the non-native speech perception and production literature. In addition, the chapter addresses the implications for non-native perception and production theories and second language speech learning and teaching.

The Appendices contain copies of written materials used in each of the experiments, such as participant information sheets and consent forms. Conference papers generated by this thesis project are also included in the Appendices: SST2018, ICPHS 2019, INTERSPEECH 2019.

Chapter 2. Non-native speech perception and production theories

2.1 Speech perception theories

Speech is the primary means of human communication. Speech perception refers to the process of detecting, categorising and discriminating the phonemes and words carried by speech signals. Native speech perception is so efficient and robust that we often take it for granted. But an examination of the underlying mechanisms of speech perception reveals just how complex a process it is. The speech stream is unlike a written sequence of words because there are no spaces between units as there are between written letters/words, referred to as *the segmentation problem*. Moreover, there is no simple or direct correspondence between phones and phonemes as perceived and the acoustic patterns generated by articulatory gestures, referred to as *the constancy problem*. There are huge variations induced by different phonetic contexts and vocal characteristics of the talkers. The fact that we can detect and sort the segments that are embedded in speech into native phonetic/phonological categories with little effort is remarkable.

Speech perception theories should address the above inherent perception problems and also consider the relation between speech perception and production as the two are intrinsically linked in speech communication. Despite the fact that the end-organ sensory receptors within our ears are stimulated by acoustic patterns including speech, there are competing meta-theoretical assumptions concerning what informational primitives, either gestural or auditory in nature, in speech perception listeners rely on to categorise these acoustic patterns into phones in a given language, and whether there is a domain-general or domain-specific mechanism for speech perception (see Table 2.1, adapted from Diehl et al., 2004).

Table 2.1 Taxonomy of major theoretical approaches to speech perception (adapted from Diehl et al., 2004).

	Special mechanism	General mechanism
Gestural	Motor theory	Direct realism
Non-gestural	Eclectic specialisations[1]	General auditory processes

2.1.1 The general auditory approach/psychoacoustic approach

The general auditory approach (as coined by Diehl et al., 2004) or psychoacoustic approach (as coined by Best, 1995) encompasses a number of theories (Diehl & Kluender, 1989; Ohala, 1996; Sussman, Fruchter, Hilbert, & Sirosh, 1998) which share that speech sounds are perceived via the same auditory and perceptual learning mechanisms involved in perceiving non-speech sounds and listeners recover messages from the acoustic signal without reference it to gestures. According to this view, the perceptual primitives in speech perception are spectral and/or temporal acoustic cues in the speech signal. Listeners extract acoustic patterns and map them to their mental representations.

This theoretical account is intuitive but is confronted with the problem of lack of acoustic-phonetic invariance in the speech signal. Two approaches have been proposed to solve this problem with different assumptions of mental representations: (1) if the mental representation is assumed to be abstract, as assumed by linguists, listeners need to filter out irrelevant information before they map

[1] Note: the lower left quadrant corresponds to claims that a special mechanism is used without gestural mediation in speech perception. This has not been developed into a coherent theory, but some infant speech researchers have indeed attributed human infants' ability to learn native phonological categories to specialised processes of auditory categorisation (Kuhl, 1991, 1993) and to an attentional or learning bias for speech sounds (Jusczyk, 1997).

the acoustic pattern to the corresponding mental representation through a process of normalisation; (2) if mental representations include all experienced exemplar details, which are stored as episodic memories, then a more complex mapping process is needed; one that does not involve normalisation processes.

Under the shared assumptions of a general mechanism for speech perception, theories and models of the general approach deal with speech perception in different ways. The *auditory enhancement hypothesis* (Diehl & Kluender, 1989) maintains that listeners are primarily sensitive to the auditory qualities of phonetic segments. They claim that many universal tendencies that have been observed across the phonemic inventories of languages arise because speech communities tend to select components that can mutually enhance auditory effects. For example, nearly all languages include the three point vowels (/i/, /a/, /u/), which are maximally dispersed in the auditory space. More interestingly, the front vowels are often produced without lip rounding whereas back vowels are produced with lip rounding. According to the auditory enhancement hypothesis, lip rounding generates a lower frequency F2 (Stevens et al., 1986) and thus enhances perception of back vowels; on the other hand, lip rounding counteracts the acoustic effects of tongue fronting, rendering rounded front vowels less distinguishable than back vowels.

The *fuzzy logical model*, another auditory theory, sees speech perception as auditory pattern recognition (Massaro & Oden, 1980) in which relevant acoustic features of a given contrast are extracted independently from the speech signal and combined according to logical integration rules. Features are assigned a probability value between zero and one capturing the extent to which a given feature is present in the stimulus. The features are then integrated to determine the degree to which the stimuli match a stored prototype. In this way, the model maps acoustic attributes onto

higher-level representations via a probabilistic process of matching features to prototypes in memory.

2.1.2 Motor theory

Motor theory (Liberman et al., 1967) was primarily motivated by the fact that a perceiver is also a speaker in human communication and thus argues that as humans we possess specialised mechanisms for both perception and production. It is based on the premise that it is economical to assume that there is one integrated process, rather than two entirely separated processes, for encoding and decoding speech.

The theory contends that the perceptual primitives in speech are not acoustic cues, but neural commands to articulators (Liberman et al., 1967), or more recently, intended gestures represented in one's mind (Liberman & Mattingly, 1985). Listeners reconstruct the intended gestures of the talker from the speech signal, which is thought to be much less susceptible to phonetic variation than general auditory processing is. In this way, the motor theory naturally links speech perception and production and resolves the problem of lack of invariants between phonemes and acoustic signals. The neural representations of speech in the motor theory are abstract, canonical linguistic representations.

In addition, motor theory argues that speech perception is processed via a specific neural module for human speech, which is supported by some phenomena posited to be speech-specific and uniquely human, such as categorical perception (Liberman et al., 1957) and the right ear advantage in dichotic listening (Studdert-Kennedy & Shankweiler, 1970). However, categorical perception has reportedly been found in perception of non-speech sounds (Miller et al., 1976) and by animals (Kuhl & Miller, 1975). Such counterevidence presents a challenge for the claim of a human-only speech-specific module for speech perception.

2.1.3 The direct realist approach

The direct realist approach (Fowler, 1989; Best, 1995) shares with the motor theory the view that perceptual targets are gestural in nature, but it argues that the actual gestures rather than the intended gestures are perceived. The gestural information correlates with acoustic patterns via the principles of acoustic physics and is directly perceived. This implies that there is no need for mental representation and consequently no need for perceptual normalisation when dealing with variability in speech as compared with a canonical mental representation. In contrast to the motor theory which argues for a domain-specific mechanism, the direct realist approach contends that speech perception mechanism is domain-general, just like perception of other events in the world. Articulatory gestures for lexical tones include laryngeal movements to raise and lower pitch (cricothyroid and arytenoid muscles) and possibly also to raise and lower the larynx itself, i.e., external muscles in the trachea. Laryngeal gestures may also result in voice quality changes such as breathiness and creakiness (Brunelle, Nguyên, & Nguyên, 2010; Erickson, 1976; Erickson, Liberman, & Niimi, 1976; Erickson & Abramson, 2013; Nguyen & Edmondson, 1998; Sagart, Hallé, Boysson-Bardies, & Arabia-Guidet, 1986). The present thesis assumes that the perceptual primitives of both native and non-native tone perception are stable assemblies of multiple laryngeal gestures.

2.1.4 Developmental considerations and model comparisons

Apart from the mechanisms underlying speech perception, another issue that all speech perception models must address is how human beings grow up acquiring the ability to perceive their native language. Infants show sensitivity to all speech segments, but as they acquire their native language, their sensitivity becomes more restricted to speech contrasts that exist in their native language and they show a decline in sensitivity to most of the contrasts that are absent in their mother tongue.

This native language attunement effect on non-native perception and production lies at the core of current cross-language speech perception and production theories and models.

For auditory/psychoacoustic approaches, acquiring a native language results in memory traces or templates and/or prototypes of native phonetic categories. Motor theory, on the other hand, holds that native phonetic input tunes the speech module that is responsible for speech perception. Direct realism maintains that children discover higher-order invariants of articulatory gestures in speech when acquiring their native language. The higher-order invariants can automatically accommodate contextual variations that are lower-order in nature, but come at a cost of losing sensitivity to the lower-order gestural invariants of non-native categories.

Non-native speech perception and production theories are built on these different meta-theoretical assumptions of perceptual targets and processes and how they are affected by language-specific experience during native language acquisition. In the next section, I outline and review the four most representative theories first.

2.2 Non-native perception and production theories

Our perception of non-native phones is constrained by our native language experience. Several theoretical models have been proposed to account for native language influence on non-native perception and production, each with different meta-theoretical assumptions about speech perception and different foci.

2.2.1 The Perceptual Assimilation Model

The Perceptual Assimilation Model (PAM: Best, 1995) was built on the meta-theoretical basis of ecological direct-realism and the principles of Articulatory Phonology (Browman & Goldstein, 1989). Thus, it assumes that the perceptual primitives of both native and non-native perception are

gestural constellations, i.e., stable assemblies of multiple gestures. To account for how attunement to the native language constrains non-native perception, PAM posits that naïve adult listeners tend to perceive non-native segments according to their similarities to or discrepancies from the native gestural constellations that are closest to them in native phonological space. These similarities and discrepancies can be indexed by the spatial proximity of constriction locations and active articulators and by similarities in constriction degree and gestural phasing, which can predict how naïve listeners perceptually assimilate the non-native phones into native categories (Best, 1995).

A non-native phone may be heard, (i) as a good to poor exemplar of a native phoneme (Categorised), (ii) as unlike any single native phoneme but within the native phonological space (UnCategorised), or (iii) as a nonspeech sound (Non-Assimilated). PAM also considers sensitivity to within category phonetic variations in perceptual assimilation. For the Categorised assimilation, the non-native phone can be perceived along a gradient from a good exemplar of that category, to an acceptable but not ideal exemplar of the category, to a deviant exemplar of the category.

Perceptual assimilation types can be predicted by comparing non-native and native articulatory similarities. For example, Best, McRoberts and Goodell (2001) tested predictions about the perceptual assimilation of three Zulu consonant contrasts by American English (AE) listeners: the voiceless versus voiced lateral fricatives, /ɬ/ vs. /ɮ/, voiceless aspirated versus ejective velar stops, /kʰ/ vs. /k’/, and the plosive and implosive bilabial stops, /b/ vs. /ɓ/. The lateral fricative contrast, /ɬ/ vs. /ɮ/, employs similar articulatory organs and constriction location to that of the AE lateral approximant /l/, but AE does not have lateral fricatives. The authors predicted that the lateral fricatives, /ɬ/ vs. /ɮ/, would be categorised as AE voiceless apical fricatives, /θ, s, ʃ/, and voiced apical fricative, /ð, z, ʒ/, or lateral, /l/, respectively because both shares the same articulators, i.e., tongue tip and dorsum, glottis, constriction locations, i.e., dental/alveolar and posterior

constrictions, and constriction degree, fricative. For the second contrast, Zulu /b/ is a short-lag unaspirated voicing plosive, similar to the allophone of AE /b/, i.e., [p]. The Zulu implosive /ɓ/ is similar to another allophone of AE /b/, i.e., [b] with full voicing. The two Zulu bilabial stops were predicted to be categorised into the single AE category /b/ because they are produced with the same organs, constriction location and degree. As for the third contrast, the voiceless aspirated velar stop /kʰ/ is almost identical to AE /k/, whereas the ejective velar stop, /k'/, has a cessation of glottal airflow during stop release, making it a deviant exemplar of AE /k/. Thus, both Zulu velar stops were predicted to be categorised as a single AE category /k/, but /kʰ/ as a good exemplar and /k'/ as a deviant exemplar of that category. The results upheld PAM predictions.

More importantly, PAM provides further predictions about discrimination of non-native contrasts according to how these contrasts are assimilated to native categories. If each non-native segment is assimilated to a different native category, the contrast forms a Two-Category assimilation. The discrimination of the Two-Category contrast is predicted to be better than if both non-native segments are assimilated into the same native category as being equally ideal, acceptable or deviant, that is, a Single-Category assimilation. Otherwise, if both non-native segments are assimilated to the same native category but differ in how they are deviant from the native "ideal", they form a Category-Goodness assimilation and the discrimination of Category-Goodness assimilation contrast is predicted to be better than that of Single-Category assimilation but poorer than that of Two-Category assimilation, i.e., Two-Category > Category-Goodness > Single-Category. Predictions derived from PAM principles have been validated against a number of studies on non-native consonant (Best, Avesani, Tyler, & Vayra, 2019; Best et al., 2001) and vowel perception (Faris et al., 2018; Tyler et al., 2014).

If one segment of the contrast is Categorised and the other is UnCategorised, the contrast forms an UnCategorised-Categorised assimilation. The discrimination is expected to be very good if the two phones fall on either side of a native phoneme boundary. Moreover, if both non-native segments are UnCategorised, forming UnCategorised-UnCategorised assimilation, discrimination will depend on their similarity to each other and to native categories.

Furthermore, Faris, Best, and Tyler (2016, 2018) have proposed three ways in which a non-native phone might be UnCategorised: (1) a *focalised* response in which the non-native phone is assimilated as primarily similar to a single L1, i.e., native language/first language, category but choices of that native phoneme fall below the defined percent choice categorisation threshold; (2) a *clustered* response in which the non-native phone is assimilated below threshold to a small set of L1 categories; or (3) a *dispersed* response in which the choice of native phone category is spread across many L1 categories, all below chance level. Native phonological influences are moderate for $\text{UnCategorised}_{\text{focalised}}$ assimilation and weak for $\text{UnCategorised}_{\text{clustered}}$ assimilation and very weak for $\text{UnCategorised}_{\text{dispersed}}$ assimilation. Consequently, a new set of predictions were proposed for the discrimination of non-native contrasts involving these UnCategorised assimilations (Faris et al., 2018). If one or both phones in the non-native contrast involve focalised/clustered assimilation and there is no overlap in the assimilated native categories, discrimination should be better due to some degree of phonological distinctiveness than if there is any perceived phonological overlap, which will further weaken phonological distinctiveness. For dispersed assimilation, if both phones of an UnCategorised-UnCategorised assimilation contrast are dispersed, discrimination accuracy will be even lower but may be slightly above chance depending on any perceived phonological overlap in categorisations. If only one phone is focalised/clustered and the other dispersed, discrimination will be good because listeners will perceive moderate to

weak L1 phonological similarity for the focalised/clustered phone and much less to no phonological similarity for the dispersed phone.

Perceived L1 phonological overlap in fact is not restricted to UnCategorised assimilation, but prevails among all types of assimilation (Antoniou et al., 2013; Best et al., 2019; Faris et al., 2018). If contrasting non-native phones are each identified with completely different sets of native categories, i.e., Two-Category/Non-overlap, UnCategorised-Categorised/Non-overlap, UnCategorised-UnCategorised/Non-overlap, the discrimination should be better than if there are one or more shared above-chance categories for the contrast assimilation, but not all choices are shared, i.e., Two-Category/Partial-overlap, UnCategorised-Categorised/Partial-overlap, UnCategorised-UnCategorised/Partial-overlap. Finally, if all the above-chance categories are the same for both non-native contrasts, the contrast is *completely overlapped*, i.e., Two-Category/Complete-overlap, UnCategorised-Categorised/Complete-overlap, UnCategorised-UnCategorised/Complete-overlap, and the discrimination should be poor or even similar to that of Single-Category. These predictions have been supported by perceptual evidence from non-native consonants (Antoniou et al., 2013), vowels (Faris et al., 2018), and in the case of Two-Category assimilation (Best et al., 2019).

In the rare and extreme cases when both non-native segments are perceived as non-speech sounds, i.e., Non-assimilable, or NA, discrimination is expected to vary from good to excellent depending on non-speech perceptual factors such as psychoacoustic salience. A telling example can be found in a study by Best, McRoberts and Sithole (1988), who tested AE listeners' perception of click consonants in Zulu, which are not phonologically contrastive in English and do not share any phonetic-articulatory features with English. AE listeners discriminated the click contrasts with high accuracy, even though these contrasts do not exist in their native language. The authors argued

that this reflected non-speech perceptual influences because the listeners perceived the clicks as non-speech sounds and hence were not constrained by L1 phonological or phonetic factors.

PAM considered both phonological and phonetic levels in predicting cross-language assimilation patterns and discrimination performance. Although naïve listeners maintain both levels in their native language, in which perceived differences at the phonetic level are systematically related to the functional linguistic categories at the phonological level, naïve listeners cannot tell which non-native phonetic distinctions comprise phonological differences in another language (Best & Tyler, 2007). What they can perceive are phonetic deviations of the non-native phones from their native phonemes. Thus, non-contrastive gradient phonetic details are included in parallel with abstract contrastive phonological features when predicting non-native assimilation and discrimination patterns from PAM principles (Best & Strange, 1992; Bohn & Best, 2012; Hallé, Best, & Levitt, 1999).

While PAM focuses more on non-native speech perception by naïve listeners, basic principles of PAM have been extended to bilingual speech perception (Antoniou et al, 2012, 2013) and production (Antoniou et al., 2010, 2011), and second language speech learning (PAM-L2: Best & Tyler, 2007).

2.2.2 The Speech Learning Model

The Speech Learning Model (SLM: Flege, 1995) assumed the psychoacoustic approach to speech perception and was originally created to account for a variety of factors that contribute to accented production by second language learners. These factors include age of learning/age of arrival (AOL/AOA: Flege & Fletcher, 1992; Flege, Takagi, & Mann, 1995; MacKay, Meador, & Flege, 2001), native language use in a second language environment (Flege & MacKay, 2004) as well as native versus-non-native language similarities/differences (Bohn & Flege, 1992; Flege, 1987). The

former two factors are important in the study of second language speech learning but are beyond the scope of this thesis. Thus, the focus here will be more on the theoretical components that are related to native language influences on non-native perception and production by naïve listeners when reviewing the model.

SLM claims that speech perception becomes attuned to contrastive phonetic elements of the native language, and phones in native and non-native languages are related perceptually at a position-sensitive allophonic level. This level of analysis is less abstract than the phonological level of Contrastive Analysis (Lado, 1957) but is still an abstract level in that indexical variations (e.g., talker variability) are not considered (Flege, 1995).

A core SLM premise is that non-native perception and production are linked. Without accurate perceptual "targets" to guide the sensorimotor learning, production of non-native phones will be inaccurate. Although not all non-native production errors are perceptually motivated, many do have a perceptual basis (Flege, 1995).

Non-native listeners may fail to perceptually discern the phonetic differences between pairs of phones in the non-native language, or between native and non-native phones. This could be caused by either (1) equivalence classification, in which distinct non-native phones are categorised to a single native category, and/or (2) native language filtering, where features of non-native phones that are important phonetically but not phonologically are filtered out.

SLM proposes three kinds of non-native categories in relation to native categories (Flege, 1992). *Identical* non-native categories are equivalent to native categories phonetically and phonologically without either acoustic or perceptual differences. They should be represented by the same IPA symbols used to represent their counterparts in the native language. *Similar* categories are represented by the same IPA symbol as the native category but with statistically significant

acoustic and audible differences. If these categories are classified as equivalent to native categories and are substituted by native categories at least initially, there will be accented production and difficulties in perception. *New* categories differ acoustically and perceptually from the most similar native categories and are represented by IPA symbols that are not used for any native phones.

SLM holds that separate phonetic categories may be established for *new* categories because they evade equivalence classification (Flege, 1991) and thus they will ultimately be better produced and perceived than *similar* categories. For example, Flege and Hillenbrand (1984) assessed the production of the French syllables /tu/ and /ty/ produced by AE and French speakers. French /u/ is considered a similar category to AE /u/, whereas French /y/ does not have a counterpart in AE. According to native speaker judgements, French /y/ vowels were produced equally well by both experienced and inexperienced AE learners, but French /u/, which is substantially more backed than AE /u/, was produced better by the experienced than inexperienced Americans, suggesting that /y/ was easier to learn than /u/. Acoustically, AE speakers produced French /u/ differently from the French talkers in terms of F2 (the second formant), but produced /y/ with F2 values equal to those of native French speakers. This suggests that a new category was formed for the new sound /y/, but not for the similar vowel /u/. This and other analogous studies (Bohn & Flege, 1992) support the hypothesis that similar non-native categories are more difficult to produce without a foreign accent than new non-native category. Nevertheless, SLM does not make explicit predictions about performance in discriminating non-native contrasts involving *new, similar*, or *identical* categories; rather, the model focuses more on the perception and production of individual non-native segments (for a discussion see Best et al., 2019).

2.2.3 The Second Language Linguistic Perception model

The Second Language Linguistic Perception model (L2LP: e.g., Escudero, 2005, 2009) accounts for second language speech learning from the initial state to ultimate attainment. For the initial state, L2LP claims that non-native listeners rely on optimal perception, a perception grammar attuned by their native language, which is comprised of cue constraint rankings based on Optimality Theory (OT; Smolensky & Prince, 1993) and its probabilistic version, Stochastic OT (Boersma, 1998). In other words, listeners will initially perceive non-native phones in line with the acoustic features of their native language, called the Full Copying hypothesis (Escudero, 2005). The model also contends that there is a direct link between perception and production (Escudero, 2005), and that perception of both native and non-native phones should match the *acoustic* properties of phones in participants' native language/dialects (Escudero & Vasiliev, 2011).

L2LP bases predictions about non-native perception by naïve listeners on the cue ranking dimensions of the native language. To illustrate, vowel duration is considered to be a non-previously-categorised dimension for Spanish listeners. Thus, the Southern British English (SBE) /i/-/ɪ/ duration difference will not be used by Spanish listeners when categorising SBE /i/ and /ɪ/. However, vowel F1 is used to distinguish L1 Spanish /i/ and /e/ and thus is considered an already-categorised dimension for Spanish L1 listeners. Since SBE /i/ and /ɪ/ both fall within the boundaries of F1 values for the Spanish /i/ category, L2LP predicts that SBE /i/ and /ɪ/ will both be categorised as Spanish /i/. This prediction is supported. In fact, empirical evidence supporting L2LP's acoustic similarity account in explaining cross-language assimilation comes from vowel perception (Escudero & Vasiliev, 2011; Escudero & Williams, 2011).

Analogous to three contrast assimilation types of PAM, but based on acoustic similarity like SLM, L2LP details three scenarios in which native and non-native contrasts can be related (van Leussen

& Escudero, 2015). A NEW[2] scenario refers to the case when the majority of productions of a non-native contrast are acoustically closest to typical or average productions of a single native segment. A SIMILAR scenario refers to when the majority of tokens of a non-native contrast are acoustically similar to the productions of two native segments. The perception and production of the non-native contrasts in a SIMILAR scenario are easier than those in a NEW scenario, because non-native listeners/speakers can use the existing native categories in the SIMILAR scenario, though they may have to adjust the boundaries, while in the case of a NEW scenario, a new non-native category needs to be created or a native category needs to be split. A third possible relation between native and non-native categories is the SUBSET scenario, c.f., UnCategorised-Categorised or UnCategorised-UnCategorised assimilation in PAM terms, where one or both non-native segments of the contrast are perceived as similar to more than one native category. The discrimination and production of this contrast is expected by the model to be relatively good.

2.2.4 The Native Language Magnet model

The Native Language Magnet (NLM, Kuhl & Iverson, 1995; Kuhl, 1994) model posits that native language experience alters perceived distances in the acoustic space underlying phonetic categories and consequently influences non-native perception and production. The model assumes that the underlying organisation of phonetic categories is the phonetic "prototype", the ideal or best instance of a category. The phonetic prototype, developed via native language experience, acts as a "perceptual magnet" for other tokens in the category. It attracts those tokens towards itself by reducing the discrimination sensitivity or the perceptual distance as compared with non-

[2] *new* in terms of individual phones in SLM is different from the NEW scenario, referring to non-native contrasts, c.f., Single-Category assimilation in PAM terms. *similar* in terms of individual phones in SLM is different from the SIMILAR scenario, referring to non-native contrasts, c f., Two-Category assimilation in PAM terms. I capitalised all letters when describing L2LP scenarios as used originally in L2LP literature.

prototypic tokens of the same category, leading to a significant asymmetry in discrimination, with good discrimination of non-prototypes from prototypes, but poor discrimination of prototypes from non-prototypes.

Evidence for the "magnet" effect comes from experiments that presented synthetic stimuli varying systematically and in equal steps along multiple acoustic dimensions. For instance, when discriminating a set of 64 2-dimensional (F1 and F2) variants of the vowel /i/, adults' and infants' showed poorer discrimination performance when the prototype of the category determined via perceptual rating served as the referent vowel as compared to when the non-prototype did (Kuhl, 1991). This effect has also been found with consonant contrasts, e.g., the velar voicing contrast /k/-/g/ (Davis & Kuhl, 1994) and the liquid contrast /ɹ/-/l/ (Iverson & Kuhl, 1996).

By extrapolation, Kuhl (1995) claims that non-native perception difficulty varies as a function of the proximity of the non-native phone to a native-language magnet. The closer it is to a native prototype, the more it will be assimilated to the native category and the more difficult will discrimination be. The native category prototypes act as magnets that filter out non-native acoustic variations. Iverson and colleagues (2003) found that American listeners attached more importance to F3 when distinguishing the /r/-/l /contrast and showed less sensitivity to acoustic differences around the phonetic prototypes, whereas Japanese listeners were more sensitive to F2 rather than F3 and had no stretching of the perceptual space in the F3 dimension. The authors attributed this cue weighting difference to native language experience and claimed that this is the reason for Japanese listeners' difficulties in /r/-/l/ discrimination. In other words, the non-native Japanese listeners failed to develop the correct prototype for non-native categories and consequently showed uniform rather than asymmetrical within-category discrimination.

NLM has contributed importantly to our understanding of non-native phones that fall within one native category analogous to Single-Category or Category-Goodness assimilations in PAM terms, but it lacks predictions for non-native contrasts that cross native phoneme boundaries, i.e., Two-Category, UnCategorised-Categorised, UnCategorised-UnCategorised in PAM terms. In addition, it is less clear whether the asymmetry between prototypes and non-prototypes reflects the magnet effect within the phonetic category or alternatively a cross-boundary categorical perception effect, as some of Kuhl and colleagues' non-prototypes have since been shown to be perceived as a different category (Sussman & Lauckner-Morano, 1995).

2.2.5 Comparing and contrasting PAM, SLM, L2LP and NLM

Meta-theoretically speaking, PAM assumes a gestural-phonetic basis for speech perception, and this differentiates it from SLM, L2LP and NLM, all of which assume the perceptual primitives to be acoustic in nature. This meta-theoretical issue aside, there are commonalities among these theoretical models. In terms of *predicting* perceptual assimilation of non-native phones into native categories, PAM, SLM and L2LP possess different approaches in assessing phonetic distance between native and non-native phones. PAM relies on gestural similarity between native and non-native phones while SLM relies more on IPA notation and acoustic analysis. Similar to SLM, L2LP uses acoustic information but with computational modelling, e.g., linear discriminant analysis. NLM does not explicitly address native/non-native phonetic distance.

Additionally, PAM, SLM and L2LP propose different ways that non-native phones can be assimilated into native phones but only PAM and L2LP make explicit predictions concerning discrimination of contrasts based on assimilation patterns. SLM focuses more on perception and production of individual non-native phones, whereas NLM is concerned more about non-native phones that are assimilated as prototypical or non-prototypical within a given native category, and

thus does not include predictions on non-native contrasts that cross native phonological boundaries.

SLM and L2LP explicitly claim that there is a perception and production link. PAM originally was proposed to solve perceptual problems, but its articulatory gesture principles can be easily extrapolated to account for production issues given that it is based on direct realistic approach to speech perception and Articulatory Phonology.

Empirical evidence supporting PAM came initially from perception of consonants (Best et al., 2001) followed by support with vowels (Faris et al., 2016, 2018; Tyler et al., 2014) and most recently, lexical tones (Chen et al., 2020; So, 2012; So & Best, 2010a, 2010b, 2011). SLM has been supported by evidence from consonant and vowel perception and production studies (Flege, 1987; Flege, 1992; Flege & MacKay, 2004; MacKay et al., 2001) whereas L2LP studies have focused more on non-native vowel perception (Alispahic et al., 2017; Elvin et al., 2014). PAM's central goal is to demystify how native language experience shapes speech perception by naïve listeners and it has been extended to account for performance of second language learners (Best & Tyler, 2007; Bundgaard-Nielsen et al., 2011a, 2011b, 2012) and bilinguals (Antoniou et al., 2010, 2011, 2012, 2013; Krebs-Lazendic & Best, 2013). Conversely, SLM considers the ultimate achievement, mostly production, of bilinguals and addresses how different factors other than native language constraints, e.g., age of learning or native language use in second language learning, contribute to accented production. L2LP claims a computational account for different stages of second language development in both perception and production.

In this thesis, PAM will be employed as the main theoretical framework for exploring native language influence on non-native perception because it is the only model that explicates both phonetic and phonological factors in native language influence and provides systematic predictions

about discrimination of non-native contrasts based on perceptual assimilation. In addition, due to its meta-theoretical basis in direct realism, its principles can be extended to predict accuracy and deviations in imitation of non-native lexical tones based on perceptual assimilation. SLM and NLM will also be compared and incorporated in the discussion of results. L2LP will not be considered further due to its focus solely on acoustic properties of non-native vowels.

Native language influences vary across different phoneme types and perception task types. To account for these variations in speech perception/imitation, it is also important to consider phonological versus phonetic *modes* of perception and how they are modulated by cognitive factors such as memory load and natural phonetic variations. The Automatic Selective Perception model (ASP: Strange, 2011) addresses these issues and is introduced in next section.

2.2.6 The Automatic Selective Perception model

The Automatic Selective Perception model (ASP: Strange, 2011) was developed to account for the online processing of continuous speech by naïve and second language listeners. The model distinguishes three types of memory components and two modes of speech perception.

Long-term memory refers to listeners' knowledge of the relationships between phonetic and phonological units. In non-native speech perception, and for naïve listeners, long-term memory contains only listeners' knowledge of the relationships between phonetic and phonological units in their native language. *Procedural memory*, also called selective perception routines in ASP, describes the knowledge listeners use to perceive L1 and L2 speech by detecting task-relevant information specified by acoustic patterns and/or specific task demands. *Short-term memory* is the rapidly fading memory trace of the fine phonetic-acoustic details of the stimuli.

According to ASP, the *phonological mode* of speech perception is accomplished via native selective perception routines which recognise the most reliable information in native phonological sequences. When non-native phones are not phonologically contrastive in the native language e.g., PAM's Single-category or Category-goodness assimilation, perception in this mode may be inaccurate, reflecting strong native language phonological influences.

On the other hand, in the *phonetic mode* of perception, the attention focus is on detecting the phonetic details, e.g., native listeners' perception of accented speech or speech produced by non-native speakers. In this mode of perception, non-native phonetic features that are absent in the native language can be detected via memory traces still held in short-term memory. Thus, the prerequisite for a phonetic mode in non-native perception is the availability of phonetic information in short-term memory.

Listeners switch between phonetic and phonological modes of perception as a function of memory load. Under low memory load, rich phonetic details are still available, and listeners tend to use a phonetic mode of perception. On the other hand, under high memory load, phonetic details have faded from short-term memory and listeners shift to a phonological mode of perception. When non-native listeners use a phonological mode of perception, they perceive stimuli according to native-language phonological categories and have difficulties discriminating the phonetic basis for non-native contrasts that are absent in their native language. When they are using a phonetic mode of perception, they show greater sensitivity to phonetic distinctions that are not used in their native language.

Werker and Logan (1985) tested the perception of the Hindi retroflex and dental stop contrast, which is absent, in AE by native AE listeners under different memory loads. Memory load was

manipulated by varying the duration of the interstimulus interval (ISI), i.e., the amount of time listeners have to keep the first stimulus in memory before hearing the second stimulus in a trial. Under low memory load (ISI = 500ms) AE listeners used a phonetic mode of perception and could distinguish the non-native contrasts that are absent in their native language, whereas under high memory load (ISI = 1500 ms) they could not distinguish the same contrast.

In addition, Asano (2017) compared the discrimination of Japanese consonant length contrasts between native Japanese versus German listeners. While both groups performed well under low memory load, native German listeners discriminated poorly under high memory load.

Given that perception precedes production in imitation, and assuming that perception and production are linked to some degree, it is reasonable to expect that non-native speakers will also switch between the two modes in an imitation task as a function of memory load. Some studies have found that non-native imitators can accurately imitate non-native phones (Rojczyk, 2012a, 2012b; Rojczyk et al., 2013) absent in their native language under low memory load, while others have shown that even under low memory load, non-native speakers have difficulties in imitating non-native phones accurately and are strongly constrained by their native language phonology (Asano & Braun, 2016; Llompart & Reinisch, 2018). Nevertheless, there is evidence that delaying imitation by inserting another task in between reduces accuracy and increases native language phonological constraints as compared to immediate imitation (Rojczyk, 2012a; Rojczyk et al., 2013), suggesting the existence of phonetic versus phonological modes of imitation.

Apart from memory load, variability of speech caused by talker or phonetic context differences makes direct mapping between detailed phonetic patterns and phonological representation difficult. Extrapolating principles from ASP, we hypothesise that high talker/vowel variability will

bias listeners to a phonological mode of perception which will activate native language phonological perceptual routines. In this mode, listeners will perceive abstract phonological features in non-native speech because the low level phonetic information is too variable and not reliable. For non-native perception, due to the activation of native phonological perceptual routines, perception will be highly constrained by native language phonological systems. On the other hand, low talker/vowel variability will bias listeners to a phonetic mode of perception because the phonetic information in the speech is consistent rather than variable. In this mode, listeners' focus will be on phonetic details. Consequently, perception of non-native contrasts, e.g., those of Category-Goodness and Single-Category assimilation, are likely to be discriminated more accurately.

Talker variability affects perception of native consonants and vowels (Nusbaum & Morin, 1992), lexical tones (Wong & Diehl, 2003) and words (Mullennix et al., 1989), as well as non-native perception (Antoniou et al., 2015; Lee et al., 2009; Magnuson & Yamada, 1994). For example, when Japanese listeners were asked to identify English words with /r/ and /l/ in word-initial position, performance was more accurate in constant than variable talkers blocks (Magnuson & Yamada, 1994). The comparative magnitude of this effect relative to native listeners varies from being comparable to being significantly different. Higher identification accuracy of Mandarin tones was found for constant-speaker tones than variable-speaker tones, but performance in high variability contexts was comparable for native and non-native English listeners (Lee et al., 2009). In an English word-monitoring task (Antoniou et al., 2015), talker variability affected both native and non-native listeners, but the effect was larger in non-native listeners and even greater in less proficient listeners, who had no immersive experience with English.

Extending this hypothesis of talker/vowel variability to imitation, it is expected that in high talker/vowel variability conditions, imitation should be phonetically less accurate and more constrained by native language phonology because native language phonological perceptual routines are activated. On the other hand, in low talker/vowel conditions, imitation should be phonetically more accurate and less constrained by native language phonological systems because they should be biased toward the phonetic mode, where the attentional focus is on phonetic details. So far as I know, there is no study on the effect of talker and vowel variability on non-native tone imitation.

In conclusion, PAM has been selected as the main theoretical model for examining native language influence on non-native assimilation, discrimination and imitation at both phonetic and phonological level in this thesis. ASP was selected in combination with PAM to provide theoretical accounts for phonological and phonetic mode of perception and imitation as modulated by cognitive factors, such as memory load and talker/vowel variability.

In the next chapter, a brief introduction of phonological and phonetic characteristics of lexical tones in the target languages chosen for the thesis research project, i.e., Thai, Mandarin and Vietnamese, is presented and followed by a review of major findings in native versus non-native tone perception and production.

Chapter 3. Perception and production of Lexical tones

3.1 Defining lexical tones and tone languages

Lexical tones refer to the use of melodic patterns, primarily pitch variations and also phonation types, as part of the phonological form of morphemes (Remijsen, 2016), the smallest units of meaning in a language. In contrast, the use of similar melodic patterns at higher levels, such as in phrases and utterances, is called intonation, which is beyond the scope of this thesis. Lexical tones can be used to contrast lexical meaning of content morphemes or words, e.g., in Mandarin /ma/ with a high level tone means *mother* whereas /ma/ with a dipping tone means *horse*, or to mark different grammatical functions, e.g., in Shilluk, a language in South Sudan, tone distinctions are used to specify different past tenses.

In a broad sense, a language that employs tones as part of the phonological specification of morphemes, or lexical items, is called a tone language. However, for some languages that use tonal differences phonologically, such as Swedish and Japanese, a large proportion of the lexicon lacks tone specification. Some linguists call these languages "pitch-accent" languages (Yip, 2002). However, there is no clear cut-off point between "pitch-accent" and "real" tone languages (Hyman, 2006). Thus, it is more appropriate to view such languages as falling along a continuum when we consider perception and production of lexical tones. Tone languages account for 70% of languages in the world (Yip, 2002) and are spread widely across Asia, Africa, America, Europe and the South Pacific (Maddieson, 2013). This thesis focuses on non-native perception and imitation of Thai lexical tones by two Thai-naïve listener groups whose native languages, Mandarin and Vietnamese, are also canonical tone languages.

Before introducing perceptual and phonetic aspects of lexical tones, it is important to first consider their phonological status. This requires examining whether lexical tones are segmental or suprasegmental, and their relationship with consonants and vowels, that is whether they are features of vowels or consonants or form a third phonological class.

Although tones can extend over multiple segments, generative phonology considers tone as segmental (Chomsky & Halle, 1968), which is also supported by phonological analysis (Duanmu, 1990, 1994; Lin, 1989)[3]. However, from the historical perspective of tonogenesis, lexical tones developed from diachronic changes in the laryngeal features of consonants (see 3.2). Studies differ in their findings about the articulation of tones relative to vowels and consonants. Some studies have found that tone gestures are coupled with onset consonant gestures rather than with the vowel nucleus (Gao, 2009), whereas others have associated tongue body position in tone production more with vowels (Shaw et al., 2016). Still others consider tones as a third phonological class, different from but interactive with vowels and consonants (Duanmu, 1990, 1994; Lin, 1989). Thus this issue is still open (see a discussion in Best, 2019)

Despite the diverse opinions on the phonological status of lexical tones, all languages use melodic patterns above word level, i.e., in phonological phrase/utterance (see Nespor & Vogel, 2007) whereas tone language speakers also use melodic patterns at or below the phonological word level

[3] I note that in later developments, and particularly when Goldsmith developed auto/suprasegmental phonology, it was argued that lexical tones in tone languages behaved independently of the vowels that they normally appeared on. The autosegemental approach provides an alternative view of the phonological status of lexical tones and can account for some phonological processes of tones in some Chinese varieties, like Wu (Selkirk & Shen, 1990), and in many Tibeto-Burman languages (e.g., Hyman & VanBik, 2004; Hyman & VanBik, 2002). However, none of this is incompatible with the argument of this thesis that for tone language speakers, lexical tones play a different role in tone languages from that of tonal units in non-tone languages, i.e., as prosodic units. The thesis focuses on cross-language perception of tones and never aims to resolve the issues of the formal phonological status of lexical tones.

(Remijsen, 2016). Thus, there is a phonological tier difference in terms of tone use between native tonal and non-tonal speakers (Best, 2019). The present thesis examined non-native tone perception and imitation by native tone language listeners, and therefore distinguishes itself from previous studies that testes non-tone language speakers, e.g., English speakers, who use melodic patterns above word level.

3.2 Tonogenesis: the birth of tones

The development of tone systems in Thai, Mandarin Chinese and Vietnamese are historically connected as these regions are close to each other geographically. This section provides a brief review of tonogenesis of these three languages based on the most frequently used model (Haudricourt, 1954), which was originally based on the analysis of Vietnamese tones.

According to Haudricourt's model[4] (see Table 3.1), an atonal or non-tonal language undergoes three stages before it becomes a full-fledged tone language. In the initial stage, i.e, stage 1 in Table 3.1, the language is atonal and syllable structures fall into four types: open syllables (CV), syllables ending with a final glottal stop (CV-ʔ), syllables ending with a final h or s (CV-h/-s), and closed (CVC) syllables with final voiceless stop consonants.

In stage 2, tones emerge: open syllables give rise to level tones, i.e., Type A as in Table 3.1; syllables ending with a final glottal stop (CV-ʔ) evolve into rising tones, i.e., Type B; syllables

[4] It should be noted that Haudricourt's model is not the only model that accounts for tonogenesis and the model didn't claim that it can account for tonogenetic paths of all tone languages. Haudricourt's model is a segmentally-driven model whereas there exist other models driven by other mechanisms such as laryngeally-based models (Thurgood, 2002, 2007), vowel-height/duration-based models (Svantesson, 1989), and aerodynamically-based models (Wayland & Guion, 2005). The present thesis does not test these models and this section of tonogenesis aims only to introduce tones in the three languages that are involved in the following experiments.

ending with a final h (CV-h) give birth to falling tones, i.e., Type C. CVC Syllables with final voiceless stop consonants remain atonal in this stage, i.e., Type D[5].

In stage 3, voiced and voiceless consonants in syllable-initial position merge and become voiceless. In order to keep the contrasts between words, the original tones are split into two registers in terms of height. The original voiceless-initial syllables evolve into high register, i.e., high tones while the voiced-initial syllables transform into low register, i.e., low tones[6]. In addition, CVC syllables with final voiceless stop consonants become tonal and also split into two registers.

Table 3.1. Tonogenesis model based on Haudricourt (1954).

Stage 1 (pre-tonal)			CV	CV -ʔ	CV -h/s	CVC
Stage 2 (tonal)			A level	B rising	C falling	D toneless
Stage 3 (split)	Voiceless C	→HIGH	A1	B1	C1	D1
	Voiced C	→LOW	A2	B2	C2	D2

[5] Alternatively, according to some laryngeally-based models (Thurgood, 2002, 2007), the three-way tonal split Haudricourt attributed to three classes of finals can be re-analysed as due to three types of laryngeal configuration: the sonorant finals led to the Type A classes, the creaky voice and final stops resulted in the Type B tone classes and the voiceless finals led to the Type C tone classes. Note, importantly, that these two accounts are *not* mutually exclusive, but could be quite compatible as they focus on two different levels of analysis (Haudricourt: phonological; Thurgood: articulatory phonetic).

[6] Again some models argued that the distinction between clear and breathy voice split each of the three earlier categories into a high-pitched and a low-pitched variant (Thurgood, 2002, 2007). Note, again, that this is an articulatory phonetic level of description, and does not necessarily carry implications about phonological level of characterisation such as we have developed in this thesis.

Since the model was originally built for explaining tonogenesis of Vietnamese tones, I will first describe Vietnamese tonogenesis, followed by that of Mandarin and Thai.

As described by the model, open syllables and syllables closed by a sonorant in Old Vietnamese became level tones as Type A in stage 2 (see Table 3.1); syllables that ended with a glottal stop evolved into rising tones as Type B presumably due to the laryngeal tension accompanying the glottal stop; syllables with final -s or -h developed into falling tones as Type C.

It is in stage 3 that the original three tones continued to split according to the voicing difference of the initial consonant. Type A tones split into two level tones: high level *ngang* and low level *huyền*; Type B tones also split into two tones with different registers and underwent additional tone changes, resulting in high rising *sắc* and low falling *nặng*; and Type C tones split into *hỏi* in the low register and *ngã* in the high register. After this, tone *hỏi* and *ngã* underwent a tonal "flip-flop" in which *hỏi* moved into high register and *ngã* into low register. In Modern Vietnamese phonology, *hỏi* behaves like a high tone while *ngã* behaves like a low tone. In syllables ending with voiceless oral stops or "checked syllables", high rising *sắc* has a variant *sắc* 2 and low falling *nặng* has a variant *nặng*2. These two tones are highly constrained by the syllable conditions, are short in duration, and differ phonetically from their parent tones (see Type D tones in Table 3.2).

Table 3.2. Tonogenesis of the Modern Northern Vietnamese tones based on Haudricourt (1954).

Stage three	CV	CV-ʔ	CV-h	CVC
Voiceless onset	ngang (A1)	sắc (B1)	hỏi (C1)	sắc2 (D1)
Voiced onset	huyền (A2)	nặng (B2)	ngã (C2)	nặng2 (D2)

The historical development of Chinese tones can also be explained by the model. According to Mei (1970), there were four syllable classes in Old Chinese, i.e., the atonal stage one: syllables ending with voiceless stops; syllables ending with vocalic or nasal segments; syllables ending with a glottal stop and syllables ending with voiced consonants.

When Old Chinese developed into Middle Chinese, tones began to emerge. According to Maspero (1912), Middle Chinese had two heights and four contours (or Si Sheng), i.e., *Ping* means level tone, *Shang* means rising tone, *Qu* means falling tone, *Ru* only applies to syllables with stop consonants in the coda, which are shorter than other tones, but the height was not used to contrast meanings. Voiceless initials, regardless of aspiration, were in the high register while voiced initials were in the low register. In other words, these contrasts were not phonological in nature but mere natural phonetic realisations. This is equivalent to stage 2 in Haudricourt's model.

From the 9th century on, the voiced initials became devoiced, making it necessary to use tones as phonologically contrasting features to maintain lexical contrasts, and leading to a system of six tones, i.e., stage 3. The four tones in Modern Mandarin derived from these six tones: the level tone or *Ping* tone split into two parts, Tone 1 or *Ying ping* and Tone 2 or *Yang ping*, which is phonetically rising and no longer a level tone; Tone 3 was inherited from *Shang* tones with voiceless initials; *Qu* tones and *Shang* tones with voiced initials turned into Tone 4. The *Ru* tones were redistributed into the other three tones.

Similarly, in Proto-Thai as reconstructed by Li (1977), there were three tones in syllables ending with sonorants and no tones for syllables ending with obstruents. After that, the three sonorant-ending tones split by syllable initial voicing conditions into six tones in parallel with Chinese and Vietnamese. After a series of tone merging processes, standard Thai evolved to have five tones. It is interesting to note that in some dialects of Thai, there are still six tones (Tingsabadh, 2001).

3.3 Characterising lexical Tones

Characterising lexical tones is important for comparing them and predicting perceptual assimilation, discrimination and imitation between languages. In this section, phonological and phonetic features of tones in Thai, Mandarin, and Vietnamese are presented and compared to pave the way for later discussion of non-native tone perception and production.

3.3.1 Phonological features of tones

Numerous phonological feature systems have been proposed for characterising lexical tones, each with its own advantages in describing tone inventories and explaining tonal phonological processes (Gruber, 1964; Maddieson, 1976; Sampson, 1969; Wang, 1967; Woo, 1969; Yip, 2001)[7]. Generally, these models have been developed mainly to consider complicated phonological processes within one language. The specification of the features is often language-dependent and thus not particularly suited to making predictions in cross-language perception studies. Such studies, however, require a more abstract and universal feature system that can characterise tone distinctions for a wide range of tone languages while at the same time minimising dependence on specific, detailed phonetic realisations. Thus, we created a parsimonious phonological feature system for this purpose, which includes perceived abstract pitch contours and heights. The contour specifications in our feature system are level, or flat, contour and the dynamic contours of rising,

[7] It has also been argued by some phonologists that tone features are unnecessary for phonological analysis, or at least not as essential as segmental features, due to the unique phonological properties of tones (Clements et al., 2011; Hyman, 2011). Nonetheless, as a parsimonious feature system was needed in this cross-linguistic perception and production research project, to characterise and compare tones phonologically across three languages, we developed a system of abstract contour and height features that would allow use to make cross-linguistic phonological level comparisons. We acknowledge that our system may not, and is not intended to, provide a perfect solution to all phonological processes across the tone systems of all tone languages.

falling and their combinations of falling-rising and rising-falling. The height specifications are high, mid, low. Contour is specified for each tone in a given language, but height is specified only when it is phonologically contrastive for tones of the same contour, e.g., height distinctions between two level, two rising, or two falling-rising tones in the same language. In the four languages involved in this thesis, height contrasts between level tones are found in Thai and Vietnamese, but the only height contrast for dynamic contours occurs in Northern Vietnamese. Mandarin tones in citation form contrast only in contour type, however, with no true minimal contrasts in phonological height.

Thai, the target stimulus language, has three level tones contrasting in height, namely, high, mid and low, and two dynamic contour tones that do not contrast in height, namely, rising and falling (Gandour, 1978). One of our three listener languages, Mandarin, has one level and three dynamic contour tones in citation form: a level tone, a rising tone, a falling-rising tone, and a falling tone (Chao, 1968).

The tone systems for the two dialects of Vietnamese are more complex. Northern Vietnamese has two phonologically level tones contrasting in height, high level *ngang* and low level *huyền*. It has four phonologically contour tones: rising *sắc*, falling *nặng*, and falling-rising tones that contrast in height: high falling-rising *hỏi*, and low falling-rising *ngã* (Nhàn, 1984). In the process of tonogenesis in Vietnamese (Haudricourt, 1954), in syllables ending with a glottal stop, that stop evolved into the rising *sắc* and falling *nặng* tones, which differ according to whether the syllable's onset consonant was voiceless versus voiced. Those two simple dynamic tones are phonologically distinct from the dynamically complex falling-rising tones *hỏi* (high) and *ngã* (low) as determined by the voiceless versus voiced initials, which instead evolved from syllables ending with -s or -h. The reason for specifying *hỏi* as high and *ngã* as low is phonologically motivated (Yip, 2002): in

tone reduplication in North Vietnamese, if the inputs are rising *sắc* or high falling-rising *hỏi*, the prefixal reduplicant surfaces as high level *ngang*, whereas if the inputs are falling *nặng* or low falling-rising *ngã*, then the prefixal reduplicant surfaces as low level *huyền* (Yip, 2002).

Southern Vietnamese has five tones in its phonological system. Four of them correspond phonologically to Northern Vietnamese: high level *ngang*, low level *huyền*, rising *sắc*, falling *nặng*. The fifth tone is reported to reflect a diachronic tone merger of the two falling-rising tones, high *hỏi* and low *ngã* (Brunelle, 2009), implying that the height distinction for falling-rising tones has been lost over time in Southern Vietnamese, which thus retains a height contrast only for level tones.

3.3.2 Phonetic characteristics of tones

There are three conventions to transcribe tone patterns: the diacritic convention used by the International Phonetic Alphabet (2015), Chao tone "letters" and Chao numbers (Chao, 1930). Chao number notation in which the pitch range of a speaker is scaled from 1 (lowest) to 5 (highest). It is used widely in studying Asian tone languages because of its flexibility in representing contour tones. For this reason, Chao numbers are used to transcribe tones in this thesis. Level tones can be represented with two identical numbers to indicate the height: 55 = extra high, 44 = high, 33 = mid, 22 = low and 11 = extra low. Two or more numbers can be used to represent contour tones: 35 = rising, 51 = falling, 45 = high rising, 241 = rising falling. All Chao transcriptions are adopted from published research and the sources are identified in the immediate context. I followed experts' opinions about the tone letters for specific tones of each language and the tone letters were not used here for readers to compare with the specific F0 patterns of the acoustic stimuli they are looking at but to provide an initial approximation, because Chao letters reflect pitch levels and contours as perceived by experts of the respective languages, rather than actual F0 values.

In this thesis, the following naming convention is used: language names plus Chao numbers. For example, three level tones in Standard Thai (T) are transcribed as: high level T45, mid level T33, low level T21 (Reid et al., 2015). It should be noted that 45 in T45 and 21 in T21 are phonetic realisations of phonologically level tones and it is quite common for extreme high/low level tones to actually start with a glide from a less extreme height (Maddieson, 1972). Thai has two contour tones: rising T315 and falling T241. Mandarin has one high level tone M55, one high rising tone M35, one dipping tone M214, and one falling tone M51 (Chao, 1968).

I examined two regional variations of Vietnamese in Chapter 5: Northern Vietnamese (NV) and Southern Vietnamese (SV). Both NV and SV have two level tones: *ngang* and *huyền*. *Ngang* is commonly considered as NV/SV44 and the other level tone *huyền* is regarded as NV/SV22 (Vũ, 1981). Both NV and SV have a rising tone, *sac*, transcribed as NV/SV35. And NV *nặng* is mid-falling and shorter in duration, noted as NV21 (Nguyen & Edmondson, 1998). SV *nặng* is phonetically falling-rising (212) but to make its Chao transcription comparable to its counterpart in Northern Vietnamese, it is also named as SV21. However, the phonetic difference from NV21 is considered when making predictions in this thesis.

The tone *ngã* in NV starts on a fall and is interrupted by glottalisation, or extremely low F0, and is thus noted as falling-rising NV415. The other tone, *hỏi*, in NV starts somewhat higher than *huyền* and drops rather abruptly, followed by a moderate rise at the end in citation form and thus transcribed as NV214 (Nhan, 1984). SV merges *hỏi* and *ngã* tones into a single tone, noted as SV214 (Brunelle, 2009).

While Chao tone letters provide a generally agreed upon and a priori phonetic characterisation of tones, a few caveats should be noted. First, Chao letters indicate perceived pitch rather than measured F0 values, as noted above. Second, Chao values are specific to a single language rather

than being universally normed across all tone languages, and therefore may be misleading when comparing tones across languages. For example, 35 in one language may not be equal to 35 in another language. Second, temporal features and the temporal location of a tone inflection are not captured by Chao numbers. Thus, in M214 it is not known if the low point 1 occurs at the temporal midpoint of the tone, or earlier or later. In other words, Chao notation does not specify whether the falling and rising portions have the same or similar durations. These temporal features are very likely relevant for perception.

For these reasons, Chao values cannot be relied on exclusively when comparing phonetic similarities/dissimilarities for cross-language perception studies. Therefore, we conducted an acoustic analysis to complement the Chao notations and provide instrumental measurements of the phonetic properties of the tones under consideration in the four languages. This laid the grounds for making predictions for the cross-language perception and imitation experiments in Chapters 5-7.

With permission, the Thai stimuli for the thesis research were adopted from an existing speech corpus (Burnham et al., 2009). Recordings were of two Thai CV syllables [ma:], [mi:] produced with all five of the Thai tones. All were meaningful morphemes in Thai. The resulting target syllables had been recorded as produced by four native Thai female speakers (M_{age} = 30.3 years), who were all born and raised in Bangkok, Thailand. These materials were recorded in citation form in a sound-treated booth at Western Sydney University using a Lavalier AKG C417 PP microphone with the sampling rate of 48 kHz and 16-bit resolution. Many repetitions were produced by the speakers, but only natural-sounding high-quality exemplars of each token were selected. Five tokens per type were selected from two participants, four from the third participant,

and six for the fourth participant for the prior research. Thus, there were 20 tokens for each of the ten Thai target items (/ma/ and /mi/ × 5 tones each, 200 Thai tokens in total).

Additional recordings were made of female native speakers ($n = 4$, for each language) of Mandarin (M_{age} = 27.0 years), Northern Vietnamese (M_{age} = 22.5 years) and Southern Vietnamese (M_{age} = 20.5 years). These informants produced the same syllables as the Thai recordings, but with each of their native tones (four tones of Mandarin elicited via Pinyin; six tones of Vietnamese elicited using orthography) eight times in random order. Thus, there were 32 tokens for each of the eight Mandarin target items (/ma/ and /mi/ × 4 tones) for a total of 256 Mandarin tokens; there were 32 tokens for each of the 12 Vietnamese target items (/ma/ and /mi/ × 6 tones) for a total of 384 Northern Vietnamese tokens. Southern Vietnamese speakers were asked to read all six tones of Vietnamese but were specifically instructed to read them in their southern accent. All were meaningful morphemes in the respective languages. These productions were recorded using a Zoom H4n digital speech recorder using the built-in mic with a sampling rate of 44.1 kHz in 16-bit stereo format within a quiet testing booth at Western Sydney University.

All of the recorded syllables were manually annotated and analysed using the Praat script *ProsodyPro* (Xu, 2013). *ProsodyPro* outputs measurements of syllable duration and 10 equidistant points of F0 values (in Hz). Raw F0 (in Hz) was normalised using the Lobanov (1971) normalisation method, and the most stable part of the tone, i.e., from 10% to 90% of the syllable, was used to calculate all F0-related measures. It should be noted that the Lobanov-normalised F0 values reflect how much an F0 value for a tone varies from the F0 mean of the speaker. The time-normalised mean F0 contours (in Figure 3.1) show general F0 patterns of Thai, Mandarin and Vietnamese tones.

Not all tone acoustic cues are equally relevant at the perceptual level: some cues, e.g., tone offsets or contours, are more relevant for tone classification than others, e.g., tone onsets. To understand acoustic feature weighting differences between languages, one could refer to studies designed specifically to examine this issue, e.g., studies indicating that Northern Vietnamese listeners relied more on voice quality than their Southern counterparts (Brunelle, 2009; Kirby, 2010), and studies indicating that Mandarin listeners were more sensitive to pitch contours than non-native listeners, e.g., French or English listeners (Gandour & Harshman, 1978; Hallé et al., 2004; Huang & Johnson, 2010).

In the present study, duration, and F0 mean, direction, extreme endpoint, and slope were selected because they have been found to correlate with perception of tones by native listeners of Thai and Yoruba (Gandour, 1978) according to a previous multidimensional scaling study. To capture these features, we calculated syllable duration, $F0_{onset}$, $F0_{offset}$, $F0_{mean}$, and $F0_{excursion}$ (maximum to minimum), which had been used to characterise level tone contrasts in a previous study (Kuang, 2013), and we added one more measure, $F0_{max\_location}$ ratio, (i.e., relative location of the F0 peak as a proportion of the duration of the tone) to distinguish differently-timed peaks in convex and concave contours (e.g. T241 and T315) (see the Appendix A, Table A1).

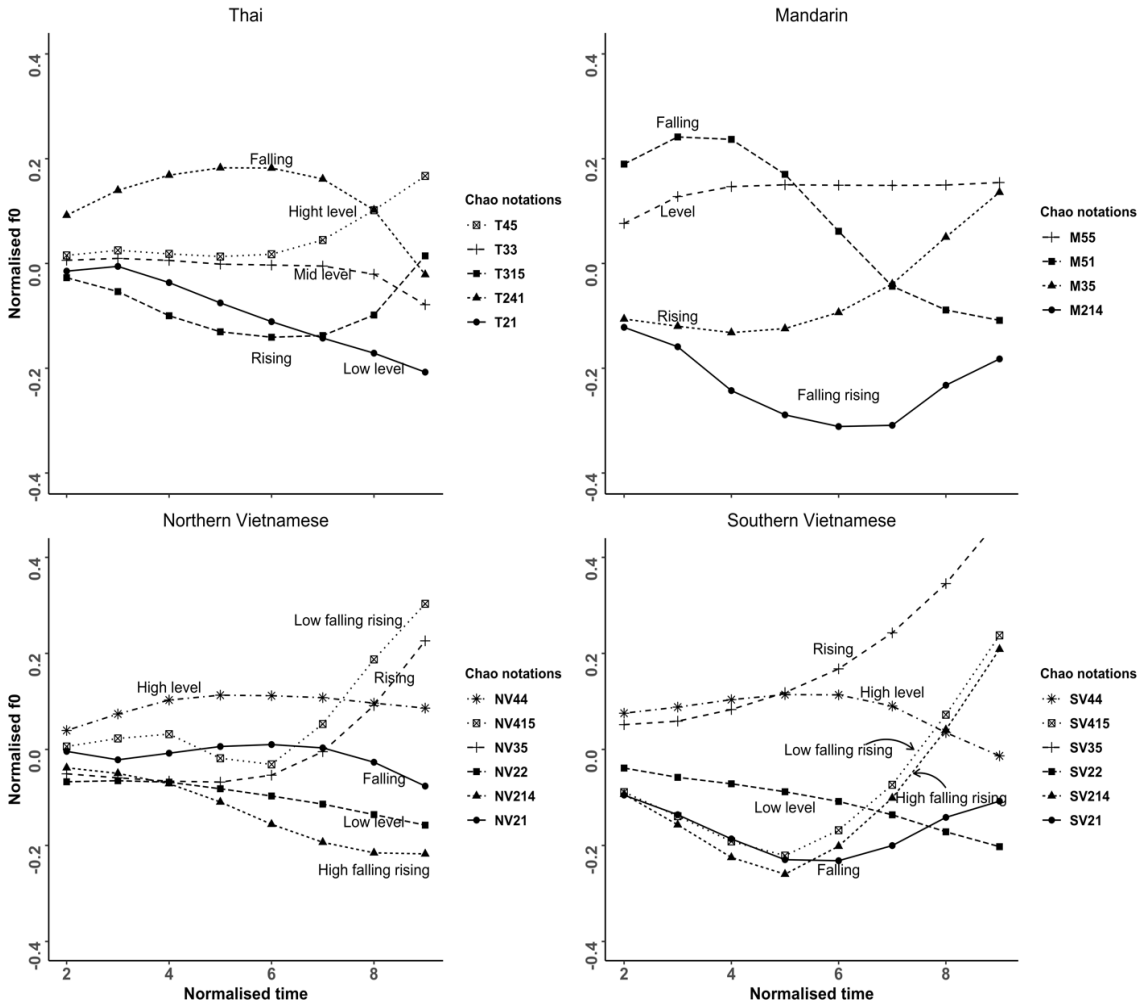


Figure 3.1. Time- and Lobanov-normalised (Lobanov, 1971) mean F0 contours of Thai (upper left), Mandarin (upper right), Northern Vietnamese (lower left) and Southern Vietnamese tones (lower right)

It is easy to see that there are both phonological and phonetic differences among tones in the four languages/dialects. For example, Mandarin has only one phonologically level whereas both Thai and Vietnamese have more than one level tone, contrasting in height. The falling tones in Mandarin and Thai are different in the contour. These phonological and phonetic differences should definitely affect non-native tone perception and imitation, which I address in this thesis.

3.4 Non-native tone perception

Pitch is primary for tone perception and the main acoustic correlate of perceived pitch is fundamental frequency or F0. Perceiving pitch in speech is different from that in non-linguistic contexts, such as in music. Perception of tones is shaped by language experience. Apart from F0 variation, other phonetic features, such as duration, intensity, and voice quality also influence tone perception. In this section, I will first review research on lexical tone perception by native speakers of Thai, Mandarin and Vietnamese. This sheds some light on the fundamental principles or mechanisms underlying tone perception and to provide the foundation for later discussion involving non-native tone perception, which is modulated by native language experience.

3.4.1 Tone perception by native listeners

Speech perception abilities are commonly examined using identification and discrimination tasks. Identification tasks require participants to listen and assign a native/non-native category label either from a closed set of response options or using orthographic symbols from that language. Discrimination tasks require listeners to distinguish between contrasting stimuli in AX or AXB/ABX tasks. Identification and discrimination can be combined in some canonical paradigms, such as that of testing categorical perception.

Most tone perception studies on languages such as Thai, Mandarin and Vietnamese, have tested monosyllabic words. Only a small number of studies have examined the perception of bi-syllabic words to investigate the nature of tone sandhi (Hsieh & Yu, 2006; Yeh & Lin, 2012). Given that the experiments in the thesis used monosyllabic words, the review below will focus on studies testing with monosyllabic words.

Both natural and synthetic stimuli have been used to explore tone perception. Naturally produced stimuli are used to identify which tones are most confusing to listeners when perceived in isolation

or in sentences. Synthetic stimuli, on the other hand, are used to manipulate acoustic dimensions of interest so as to determine their role in tone perception.

Naturally produced monosyllabic Thai tones produced by one male speaker were identified perfectly by Thai native listeners (Abramson, 1962). However, when taker variability was introduced (Abramson, 1975, 1976), the accuracy of identification of the mid and low tones decreased. Similarly, native Mandarin speakers perceive naturally produced Mandarin tones with high accuracy, although in some cases M35 and M214 confused listeners (Chuang & Hiki, 1972), apparently because M35 and M214 have similar onset F0 values and M214 turns into M35 when the tone sandhi rules apply.

The identification of Vietnamese tones was reported to vary across native listeners of different dialects. Vũ (1981) tested identification of Northern, Central and Southern Vietnamese tones by native listeners. For natural stimuli, listeners of three dialects showed overall near perfect performance in identifying tones from all three dialects in meaningful contexts, suggesting that these dialects are mutually intelligible. However, when listeners were asked to identify tones in isolated syllables, performance varied as a function of the dialect of the tones. Listeners on their performed better on their native tones than did listeners from the other dialects, and in turn performed better on native tones than on other-dialect tones. As for the performance on specific tones, Southern Vietnamese listeners showed poor identification of *hỏi* and *ngã* produced in Northern Vietnamese (Brunelle, 2009; Vũ, 1981) and production of this tone contrast in Southern Vietnamese was not distinguishable for Northern Vietnamese listeners (Vũ, 1981).

Northern Vietnamese and Southern Vietnamese native listeners employed different acoustic cues in identifying synthetic Northern Vietnamese tones in isolation. Brunelle (2009) imposed 41 F0 contours, i.e., 6 level, 30 simple rising and falling, 5 complex falling- rising or rising-falling

contours on three Northern Vietnamese tones of different phonation types, *ngang* with modal voice), *hỏi* with medial creaky voice and *nặng* with strong final glottalisation. Northern Vietnamese listeners attached more importance to voice quality than Southern Vietnamese listeners.

Another important question in native tone perception is whether F0 alone is sufficient and/or necessary for identifying lexical tones. The answer varies across different languages. Native Thai speakers can easily identify syllables superimposed with synthetic F0 contours as lexical tones (Abramson, 1962, 1975). However, when F0 information is not available, as in the whispered Thai tones, listeners are not able to identify different tones (Abramson, 1972). This finding suggested that F0 information is necessary for tone identification for Thai native speakers.

In order to determine whether F0 is sufficient for tone identification, Abramson (1978) used synthetic level tones of 16 F0 levels from 152 to 92 Hz and asked Thai native listeners to choose from three native level tones. There were very high identification peaks[8] for the synthetic low tone (90%), mid tone (73%) and high tone (88%). When movements were added to the F0 contour, going from 120 Hz at the onset, but ending at 16 different points ranging from 152 Hz to 92 Hz, the identification rate went up from 88% to 94% for the high level tone and from 73% to 84% for mid level tones, but were unchanged for the low level tone. These findings show that Thai level tones are not perceived as perfectly level but rather involve some movement, and that F0 is sufficient for native listeners to identify Thai tones in citation form.

[8] The baseline as in identification of naturally produced Thai tones in Abramson (1975) is 96.6% for low tones, 97.9% for mid tones, and 100% for high tones.

Similar to Thai tones, F0 is both necessary and sufficient for Mandarin tone identification, and identification is poor when F0 was suppressed (Howie, 1972). However, unlike findings in whispered Thai, Mandarin native listeners were able to use secondary acoustic cues when F0 information was not available (Liu & Samuel, 2004).

Perception of Vietnamese tones differs from what has been found with Thai or Mandarin in that F0 alone is not sufficient for identifying all Vietnamese tones (Brunelle, 2009; Vũ, 1981). When only F0 information was available, Vietnamese listeners showed poor identification even when the tones were from their own dialect (Vũ, 1981). There are dialect variations in perceptual cues for Vietnamese tone perception: Northern Vietnamese tones are cued by a combination of pitch and voice quality, whereas perception of Southern Vietnamese tones is purely pitch based (Kirby, 2010).

3.4.2 Tone perception by non-native listeners

Non-native listeners have problems categorising and discriminating non-native tones, but not all non-native tones are equally difficult. Perceptual difficulties may vary as a function of listeners' first language backgrounds, as previously reviewed for non-native vowels and consonants. However, a crucial difference between non-native tone and vowel/consonant perception is that all languages have vowels and consonants but not all have lexical tones. Thus, native language influences can be different for these two groups of listeners. In other words, there is a theoretical distinction between non-tonal language non-native listeners, whose native language influence is at the phrasal or utterance level, and tonal language listeners whose native language influence is at the segmental phonological level (Best, 2019).

Earlier studies on non-native tone perception started with non-tonal language learners of Mandarin. For example, Kiriloff (1969) found that English-speaking learners of Mandarin often confused

M35 and M214 in perception. In recent years, an increasing number of studies have investigated perception of tone languages other than Mandarin, including Cantonese (Wu, Bundgaard-Nielsen, Baker, Best, & Fletcher, 2015), Thai (Burnham et al., 2015; Schwanhausser & Burnham, 2005; Wu et al., 2014) or Toura (a Niger-Congo language with four level tones, Chiao, Kabak, & Braun, 2011), by non-native listeners from a variety of languages other than English, including non-tonal/quasi-tonal[9] Korean (Tsukada & Han, 2019), non-tone, non-stress languages such as French (Hallé, Chang, & Best, 2004; So & Best, 2011, 2014), the pitch accent languages Swedish (Burnham et al., 2015) and Japanese (So & Best, 2010b) and tone language listeners. e.g., Cantonese (So & Best, 2010a), Vietnamese (Chiao et al., 2011; Tsukada, 2019), Thai (Tsukada, 2019; Wu et al., 2014), Burmese (Tsukada & Kondo, 2018) and Hmong (Wang, 2013).

Discrimination difficulties vary across language backgrounds. Tsukada and colleagues conducted a series of studies with non-native listeners of not only non-tonal languages such as English (Australia), Korean (Tsukada & Han, 2019), but also pitch-accent languages such as Japanese (Tsukada et al., 2016) and tonal languages such as Burmese (Tsukada & Kondo, 2018), Thai and Vietnamese (Tsukada, 2019). Since these studies used the same set of Mandarin tone stimuli and used A-prime (A') scores (Snodgrass et al., 1985) to indicate discrimination accuracy, it is possible to compare discrimination performance across different language groups (see Table 3.3). Apart from the M214-M51 contrast, which was discriminated well by all non-native listener groups, there were huge variations in discrimination accuracy of other contrasts across the different language backgrounds.

[9] It should be noted that most dialects of Korean are now quasi-tonal (Bang et al., 2018; Kang, 2014; Kang & Han, 2013; Silva, 2006).

To examine native language influences, i.e., why some tones were well discriminated by one language group but not another, it is essential to compare native and non-native lexical tones at both phonological and phonetic levels and/or to consider perceived similarity between native categories and non-native tones, i.e., perceptual assimilation patterns via a cross-language categorisation task. Unfortunately, the series of studies by Tsukada and colleagues did not include cross-language categorisation tasks. Thus, it is not possible to explain the variations in discrimination accuracy across different language backgrounds, especially for non-tonal language listeners.

Table 3.3 Mean discrimination scores (A') of Mandarin tone contrasts by non-native listeners. Data from English (Australia), Korean (Tsukada & Han, 2019), Japanese (Tsukada et al., 2016), Burmese (Tsukada & Kondo, 2018), Thai and Vietnamese (Tsukada, 2019). Standard deviations are in parentheses. An A' score of 1 indicated perfect sensitivity, whereas an A' score of 0.5 or lower indicated a lack of sensitivity (Snodgrass et al. 1985).

Tones	Burmese	Thai	Vietnamese	Japanese	Korean	English	Mandarin
T1-T2	0.51 (0.23)	0.75 (0.17)	0.73 (0.18)	0.73 (0.16)	0.74 (0.24)	0.62 (0.19)	0.99 (0.01)
T1-T3	0.63 (0.22)	0.84 (0.09)	0.87 (0.05)	0.81 (0.16)	0.85 (0.13)	0.83 (0.12)	0.98 (0.01)
T1-T4	0.67 (0.19)	0.80 (0.12)	0.62 (0.17)	0.80 (0.12)	0.71 (0.21)	0.65 (0.16)	0.98 (0.01)
T2-T3	0.47 (0.14)	0.62 (0.13)	0.76 (0.11)	0.60 (0.15)	0.59 (0.23)	0.66 (0.11)	0.97 (0.01)
T2-T4	0.80 (0.12)	0.93 (0.05)	0.84 (0.20)	0.87 (0.11)	0.85 (0.19)	0.66 (0.22)	0.99 (0.01)
T3-T4	0.80 (0.13)	0.95 (0.05)	0.92 (0.02)	0.90 (0.08)	0.91 (0.09)	0.89 (0.07)	0.98 (0.01)

Several studies have explored native language influences on non-native tone perception by extending principles of PAM to lexical tone perception. PAM predicts discrimination performance based on perceptual assimilation patterns (see Chapter 2, section 2.2.1).

For example, So and Best (2014) examined the perceptual assimilation of Mandarin tones in sentence context into native prosodic categories by English and French listeners who were naïve to Mandarin. English distinguishes between stressed versus unstressed syllables using pitch, loudness, duration and vowel quality, whereas French does not employ lexical-level stress and uses pitch variations only at a supra-lexical level.

Results showed that English and French listeners were able to assimilate Mandarin tones into their native prosodic categories, but they differed in their assimilation patterns, which could be attributed to the used of lexical stress in the language. M55 was Categorised as statement by English listeners but UnCategorised by French listeners. M35 was split by both language groups across question and statement, though with different percent choice and ratings, and thus UnCategorised. M214 was Categorised by both language groups as a statement but with different percent choice and category-goodness ratings. M51 was Categorised by both language groups but into different categories, as a statement for English listeners and exclamation for French listeners.

The authors also extrapolated PAM principles to predict discrimination performance based on the assimilation findings (So & Best, 2014). M214-M51, which was assimilated as Two-Category assimilation was better discriminated by French listeners than M55-M51 and M35-M214, which were assimilated as UnCategorised-Categorised/Complete-overlap, in native response categories as predicted. However, M55-M51 was assimilated by the English listeners as Single-Category but was more difficult to discriminate than were M55-M214 and M214-M51, which were both also assimilated as Single-Category. Similarly, M35-M214 assimilated as UnCategorised-Categorised/Complete-overlap contrast by the English listeners was more difficult to discriminate than M55-M35 and M35-M51, also assimilated as the same type. M55-M51 and M35-M214 were

discriminated worse than M55-M214 and M35-M51 by French listeners, although they assimilated all of these tone contrasts as the same type, UnCategorised-Categorised/Complete-overlap.

The fact that contrasts of the same assimilation types were discriminated with different accuracy levels may be accounted for by considering phonetic factors in assimilation, as reflected in percent choice and goodness ratings. In other words, even though the contrasts fall into the same assimilation type at the phonological level, they may differ in listeners' residual phonetic sensitivity to the deviation of each tone from the corresponding native tone. However, the authors raised this possibility but did not consider it systematically in their study.

Although naïve listeners of non-tonal languages can assimilate non-native tones into prosodic categories, the general strength of categorisation is relatively low (So & Best, 2014). As mentioned earlier, for these listeners, the influence of the native language comes from a different tier of the prosodic hierarchy, i.e., from the phrasal or utterance level. This may contribute to the weak categorisation and relatively poor predictions for discrimination performance. However, non-native tone perception by tone language listeners could be qualitatively different from that of native non-tonal listeners (Best, 2019). For example, Wu and colleagues (2015) assessed perceptual assimilation of Cantonese tones by Mandarin listeners and tested predictions about discrimination based on the assimilation results. Category-Goodness contrasts were discriminated more poorly than Two-Category contrasts as predicted. For the UnCategorised-Categorised contrasts, overlap in native response categories reduced discrimination accuracy, in line with PAM principles.

In addition, Reid and colleagues (2015) investigated perceptual assimilation of Thai tones by Mandarin and Cantonese listeners. Mandarin listeners discriminated Two-Category contrasts better than Category-Goodness contrasts, which in turn were better discriminated than Single-Category pairs, as predicted. For the Cantonese group, however, the discrimination of Two-

Category and Category-Goodness pairs did not significantly differ. The variations in prediction accuracy reflect complexity in non-native tone perception.

The above reviewed studies generally supported the extension of PAM principles to explain native language influences on non-native tone perception and to predict discrimination based on assimilation patterns. However, these studies failed to systematically disentangle native language phonological and phonetic factors in affecting assimilation and discrimination. Chapter 5 examines how native tone language phonological and phonetic factors affect non-native tone assimilation as reflected in assimilation types and percent choice/goodness ratings, respectively. In Chapter 6, when predicting non-native tone discrimination, I consider both native phonological effects as indicated by assimilation type and phonological overlap, as well as native phonetic effects as indicated by indices that combine percent choice and goodness ratings, to account for discrimination performance differences between contrasts of the same assimilation type.

3.5 Non-native tone production

The present thesis also examines native language phonological versus phonetic influences in imitation of Thai tones by Mandarin and Vietnamese participants and compares it with their perception of the same target tones.

Most previous studies on non-native tone production tested non-tone language speakers, such as English speakers (Wang et al., 1999). For example, Wayland (1997) tested production of non-native Thai tones by native English speakers. Tone production by English speakers deviated significantly from that of native Thai speakers in F0 valley, i.e., the lowest F0, for all five tones. In addition, according to native Thai speakers' judgements of accentedness, non-native

productions of level Thai tones, i.e., T45, T21, T33, were more accented than contour tones, i.e., T241, T315.

In addition, Wang and colleagues (2003) examined the production of four Mandarin tones by native English learners of Mandarin at the beginner level. The most problematic tone for English speakers to produce, both before and after perceptual training, was M214 as judged by native Mandarin speakers. Acoustically speaking, M55 was produced with native-like patterns before and after perceptual training but M35 and M214 productions were very deviant from the native tones. The rising F0 contour of M35 before training started late and did not reach as high as the native norm while the rising contour of M214 started earlier than native contours and the turning point was not as low. Thus, the distinction between M35 and M214 as produced by English speakers was blurred, resulting in perceptual confusion for the native Mandarin judges. M51 was also produced with a lower initial F0 and less steep contour than the native norm. Thus, the falling tone was not fully realised in non-native production. To our knowledge, no study has investigated non-native production of Vietnamese.

These studies revealed difficulties in non-native tone production but failed to show any connection between perception and production. Perception data in a perceptual training study (Wang et al., 2003) on English speakers were compared to production data to elucidate the relationship between perception and production. They found a high correlation between error patterns in perception and those in native speakers' judgement of non-native production. Nevertheless, there were differences in the direction of confusion between perception and production. M35 was incorrectly perceived as M214 more often than the reverse direction, but M214 was incorrectly produced as M35 more often than the reverse. Similarly, M51 was incorrectly perceived as M214 more often than the reverse, but M214 was produced as M51 more often than the reverse. These asymmetries indicated

the complex relationship between non-native tone perception and production, which remains unresolved. Since English is a non-tonal language, however, it is impossible to disentangle native phonological versus phonetic factors in the correlation between perception and production.

Imitation combines perception and production in a natural way, and is among the first things language learners must do when learning a foreign language. The imitation of Mandarin tones by English speakers was more accurate than their identification and production of Mandarin tones (Hao & de Jong 2016), suggesting that imitation can bypass some aspect of native phonological constraints. However, since English is a non-tone language, English imitators were not affected by their native language at a phonological level. It therefore remains unresolved how native language phonological and phonetic factors impact non-native tone imitation.

In another study, native speakers of Cantonese, a tonal variety of Chinese languages, who had learned Mandarin as an L2 were asked to identify, imitate and read Mandarin tones (Hao, 2012). The correlation between their tone identification and imitation was not significant. Cantonese listeners performed better in imitating than identifying and reading Mandarin tones. This again suggests that imitation does indeed bypass some aspects of native phonological constraints, in this case even for participants who speak another tone language. Nevertheless, as verified by native Mandarin speakers, the Mandarin falling rising tone was imitated poorly by Cantonese imitators, who failed to accurately realise the final rise. This was as expected because the participants Categorised Mandarin falling rising tone as the Cantonese low falling tone, which does not have a final rise.

However, Cantonese speakers varied in degrees of exposure to Mandarin, rather than being completely Mandarin-naïve. Their variation in Mandarin proficiency may have confounded the results. Low proficiency learners and naïve participants would be expected to be affected more by

their native language than high proficiency learners in imitation. Thus, it would be desirable to test imitation of a different tone language by native tone language imitators who are naïve to the target language. The imitation experiments in the thesis contribute to our understanding by examining non-native imitation of Thai tones by Mandarin and Vietnamese imitators.

In summary, Thai, Mandarin and Vietnamese have lexical tone systems differing in many aspects such as the number of tones, phonological features and phonetic characteristics. These differences are expected to affect how Mandarin and Vietnamese listeners perceive and imitate Thai tones. Very few studies have explored non-native perception and imitation of Thai tones by tone language listeners. None of previous non-native tone perception and imitation studies have examined how cognitive factors could bias listeners toward phonological and phonetic mode of perception. In the next chapter, I will summarise major contributions that this thesis makes and introduce the three experimental chapters that follow it.

Chapter 4. The research niches and experiment series

4.1 Research niches

A central aim of this thesis is to examine native phonological and phonetic factors in non-native tone perception and imitation. A second aim is to investigate how the cognitive factors of memory load, talker and vowel variability modulate non-native tone perception and imitation by shifting listeners between phonological and phonetic modes of perception, in which native phonological constraints and residual phonetic sensitivity function differently.

First, most studies examining native language influences in non-native perception have focused on consonants and vowels. Non-native tone perception studies testing non-tonal language listeners cannot address the issue of native phonological influences because for these listeners tones function at a mismatched prosodic level rather than the segmental phonological level. Among the prior studies that tested non-native perception with tone language listeners, their native languages were limited, e.g., Mandarin or Cantonese. In addition, more often than not those studies only tested categorisation or discrimination, but not both, making it difficult to evaluate predictions from non-native perception theories. A forced-choice identification task with goodness ratings was used in chapter 5 and 6 for evaluating perceptual assimilation. As some researchers have pointed out, in a forced choice categorisation task if a stimulus is dramatically different from native tones but still closer to one specific tone than to others, it will still be Categorised. But in this case, goodness rating will be low, distinguishing it from the case where a non-native tone is similar to a native tone and consequently Categorised. I chose the present task because it is commonly used in almost all research that investigate native language influence regardless of the theoretical assumptions, be it PAM or SLM or L2LP and other tasks, such as asking participants to write response, can also bring issues in term of interpreting the results.

Second, most previous studies on non-native tone perception and production have tended to treat them as being constant across tasks and contexts without considering the contribution of cognitive factors (Strange, 2011). Under low memory load and in constant talker or vowel conditions, naïve listeners can use a phonetic mode of perception and can detect phonetic details (Werker & Tees, 1984). But when memory load and stimulus variability increase, listeners shift to a phonological mode of perception and native language phonological constraints become stronger due to the activation of native phonological perceptual routines (Strange, 2011; Strange & Shafer, 2008). Communication in the real world requires listeners to perceive speech under varying conditions where different modes of perception are activated. Therefore, taking cognitive factors into consideration will offer a more ecologically valid picture of non-native speech perception and production.

Third, very few psycholinguistic studies on lexical tones have directly examined the relationship between non-native tone perception and production. However, there is a need to bridge perception and production research due to the importance of the relationship between the two in perception theories, communication and language learning. On one hand, in everyday communication speakers and listeners keep changing their roles, so there must be an efficient link between speech perception and production. On the other hand, in second language speech learning, the perception of L2 phonological structures appears to be a major determinant of L2 accentedness in production (Flege, 1995). Both perception and production are indispensable parts of learning a new language. Without accurate perceptual "targets" to guide the sensorimotor learning of L2 phonemes, production of the L2 phonemes will be inaccurate (Flege, 1995). All of these observations underline the need for including perception and production/imitation within a single coherent project. This allows us to relate imitation performance with perceptual assimilation results in order

to determine how native language phonological and phonetic factors affect non-native tone production.

By addressing the above issues, this thesis aims to broaden the understanding of non-native tone perception and imitation. Thai tones were used as the target language for Mandarin listeners and less often studied Vietnamese listeners. Cognitive factors of memory load, talker and vowel variability were systematically manipulated in both perception and imitation experiments and were expected to shift listeners' mode of perception and imitation between the phonological and phonetic levels.

4.2 The experimental series

There are three series of experiments in the thesis as reported in Chapters 5-7. The experimental chapters are written in the form of journal articles, and all have been submitted to journals, with the first experimental chapter now being in press. A brief introduction of each Chapter is provided below.

4.2.1 Chapter 5: Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai tones by Mandarin and Vietnamese listeners

The first experimental chapter reported in the thesis examines phonological and phonetic factors in Mandarin, Northern and Southern Vietnamese listeners' assimilation patterns for the lexical tones of another tone language, Thai. Hypotheses were derived from the Perceptual Assimilation Model (PAM: Best, 1995) principles, which consider both native phonological effects, i.e., Categorised vs UnCategorised assimilation types and phonetic factors, i.e., percent choice and category-goodness ratings.

4.2.2 Chapter 6: Cognitive factors in the perception of non-native tones by tone language listeners

The second experimental chapter[10] reports a second series of perceptual experiments that examined how perceptual attunement to phonological and phonetic properties of the native language affected non-native tone perceptual assimilation and discrimination by tone language listeners. Cognitive factors were manipulated to bias listeners toward a phonological versus phonetic mode of perception in which native phonological constraints and residual phonetic sensitivity functions differently.

In separate experiments, native Mandarin and Vietnamese listeners discriminated five pre-selected Thai tone contrasts based on assimilation results from Chapter 5. These contrasts were predicted by the Perceptual Assimilation Model (PAM, Best, 1995) to be discriminated differently. Then the listeners assimilated the five Thai lexical tones into their native tone categories. Both discrimination and perceptual assimilation experiments systematically manipulated the cognitive factors of memory load, talker and vowel variability. Memory load was manipulated by prolonging the time between two stimuli or ISI in the discrimination experiment or delaying the signal for response in the categorisation experiment. Talker and vowel variability of the stimuli were systematically blocked or mixed in blocks.

4.2.3 Chapter 7: Cognitive factors in the imitation of non-native tones by tone language listeners

The third experimental chapter presents a series of non-native tone imitation experiments with cognitive factor manipulations similar to the perceptual experiments in Chapter 6. It examined

[10] It should be noted that participants in Chapter 6 and 7 are the same but different from those in Chapter 5. In Chapter 6 and 7, all Vietnamese participants spoke the Southern dialect.

how native phonological and phonetic factors in non-native speech perception (Perceptual Assimilation Model [PAM]: Best, 1995) affect non-native imitation of Thai tones and how cognitive factors (memory load, talker and vowel variability) bias toward phonological versus phonetic selective perception effects in imitation. Thai-naïve native Mandarin and Vietnamese participants imitated the five Thai tones under high versus low memory load, in which stimulus talker and vowel were either constant or variable within blocks.

Chapter 5. Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners

Juqiang Chen[a*], Catherine T. Best[a,b*], Mark Antoniou[a]

[a] Western Sydney University, The MARCS Institute for Brain Behaviour and Development, Penrith, NSW 2751, Australia

[b]Haskins Laboratories, New Haven CT, USA

Accepted: Journal of phonetics

Chen, J., Best, C. T., & Antoniou, M. (2020). Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners. *Journal of Phonetics*, *83*, 101013. https://doi.org/10.1016/j.wocn.2020.101013

5.1 Introduction

People form native phonological and phonetic categories through their experience hearing and speaking that language. This facilitates perception of the native language, but it reduces sensitivity to non-native phonemes/phones. Adults have difficulties in categorising and discriminating non-native consonants and vowels (Polka, 1992; Polka, 1995; Strange, Akahane-Yamada, Kubo, Trent, & Nishi, 2001). However, the degree of difficulty can vary greatly, depending on the phonological and phonetic similarities and differences between the listeners' native (L1) language and the non-native language. Naïve listeners are unable to discern which non-native phonetic distinctions are phonologically contrastive and thus cannot distinguish between phonological and phonetic levels in the non-native stimuli (Best & Tyler, 2007). But they can access both phonological distinctions and phonetic details in their own language as a reference framework for categorising unfamiliar non-native contrasts. Thus, native perceptual constraints can operate at either an abstract phonological level or a lower phonetic level in perception of non-native speech, or a combination of the two, to affect naïve listeners' responses to non-native phones. So far most studies on this issue have only examined consonants (Best & Strange, 1992; Bohn & Best, 2012; Hallé et al., 1999) and vowels (Faris et al., 2018; Tyler et al., 2014).

More than 60% of the languages in the world are tone languages, in which pitch variations change the lexical meanings of words (Yip, 2002). However, relatively little is known about how native phonological and phonetic factors interplay in perceptual assimilation of non-native lexical tones. This question has not been resolved by previous cross-language assimilation studies with listeners of non-tone languages, such as English or French (So & Best, 2014). Such comparisons can reveal how non-tone language listeners perceive a type of segmental-level phonological element that does not exist in their phonological systems, unlike consonants and vowels which exist in all languages

(Best, 2019). In other words, non-tone language listeners may assimilate non-native tones into their native intonation categories, i.e. into prosodic, rather than segmental categories. However, that type of assimilation is likely to differ qualitatively from assimilation to a corresponding native tone category by listeners of other tone languages (Best, 2019). Since there are no functional lexical tone categories at the segmental level in the L1s of non-tone language listeners, studies of their perception of non-native tones cannot reveal how L1 segmental-tier phonological and phonetic factors affect non-native lexical tone perception.

Only a limited number of studies have actually tested non-native tone perceptual assimilation by tone language listeners, and they show inconsistent assimilation patterns (Reid et al., 2015; Wu et al., 2014). The variations in perceptual assimilation reflect the complexity of tone perception in general and entail further investigation of the phonological versus phonetic basis of non-native tone assimilation, which have not been systematically examined in previous studies of non-native tone perception. It is important, as well, to include listeners of additional tone languages beyond Mandarin and Cantonese which are both spoken in China. The current study addresses this issue by examining perceptual assimilation of Thai tones by three listener groups, Mandarin, Northern Vietnamese and Southern Vietnamese. The phonological and phonetic differences and similarities among the target language and the listener languages provide an ideal ground for testing native tone language influences from both the phonological and phonetic levels of characterising lexical tones.

5.1.1 Phonological and phonetic influences on cross-language assimilation

To account for L1 influences on non-native speech perception, several theoretical models have been proposed: the Perceptual Assimilation Model (PAM, Best, 1995), the Native Language Magnet model (NLM, Kuhl & Iverson, 1995), the Speech Learning Model (SLM, Flege, 1995)

and the Second Language Linguistic Perception model (L2LP: e.g., Escudero, 2005; Escudero & Vasiliev, 2011; Escudero & Williams, 2011). The focus of the current study is on adult tone-language listeners' perceptual assimilation of naturally produced non-native tones. NLM has its focus on within-category discrimination of synthetic vowels and consonants, primarily by infants. SLM has concentrated more on second language learners and their second language speech production, rather than primarily addressing perceptual assimilation (or perceived similarity) by naïve listeners, the interest of the current study. L2LP has focused mainly on second language learners' perception of vowels with predictions based only on acoustic similarities, not considering abstract phonological features. Thus, these models are not ideal to be employed as the main theoretical framework for exploring native language phonological and phonetic influences on non-native perceptual assimilation of lexical tones.

PAM is the model best suited to address the questions of the current study on non-native lexical tone assimilations by naïve adult listeners of other tone languages because it considers native language influence at both phonological and phonetic levels. At the phonetic level, PAM assumes that listeners perceive the articulatory-phonetic information carried in speech, positing that naïve adult listeners perceptually assimilate non-native phones into native categories based on their perceived articulatory-phonetic (articulatory gestures) similarities to native phonemes. Articulatory gestures for lexical tones would involve laryngeal movements to raise and lower pitch (cricothyroid and arytenoid muscles) and possibly also to raise and lower the larynx itself, i.e., external muscles in the trachea. Laryngeal gestures may also result in voice quality changes such as breathiness and creakiness (Brunelle, Nguyên, & Nguyên, 2010; Erickson, 1976; Erickson, Liberman, & Niimi, 1976; Erickson & Abramson, 2013; Nguyen & Edmondson, 1998; Sagart, Hallé, Boysson-Bardies, & Arabia-Guidet, 1986). However, extrapolating from PAM principles

at the phonological level, we assume here that tone features are abstractions from articulatory-phonetic information. Therefore, in the present study we do not describe tone features in articulatory terms but rather as more abstract specifications. PAM will be employed as the main theoretical framework while NLM and SLM will also be compared and incorporated in the discussion of results where relevant.

Within the PAM framework, depending on phonological and phonetic similarities between a given non-native phone and the listener's native phonemes, they will perceptually assimilate it in one of the following ways: (i) as an exemplar of a native phoneme (Categorised) but will also perceive its phonetic goodness of fit to that native category (ranging from excellent to poor fit); or (ii) as unlike any single native phoneme but within the native phonological space, i.e., perceived as speech (UnCategorised), or (iii) as a nonspeech sound (Non-Assimilable). Faris, Best, and Tyler (2016, 2018) have extended PAM principles to account for three different ways in which an UnCategorised non-native phone can be perceived: (1) a *focalised* response in which the non-native phone is assimilated as primarily similar to a single L1 category but choices of that native phoneme fall below the defined categorisation threshold; (2) a *clustered* response in which the non-native phone is assimilated below threshold to a small set of L1 categories; or (3) a *dispersed* response in which the assimilation of a non-native phone category is spread across many L1 categories (all below chance level). In this way, PAM provides a systematic description of perceptual assimilation of non-native tones that includes both Categorised and UnCategorised responses[11], and allows for predictions about non-native perceptual assimilations based on both phonological and phonetic similarities to native phonemes.

[11] We do not expect non-native tones to be assimilated by tone language listeners as nonspeech, i.e., as Non-Assimilable.

According to PAM principles, native language phonology constrains perceptual assimilation of non-native phones to native categories. For example, if a non-native phonological contrast does not exist in the native language, naïve listeners may assimilate both members of the non-native contrast to a single native phonological category (Single-Category assimilation) because they are not sufficiently sensitive to the non-native phonetic distinction. Strong native phonological effects on Single-Category assimilation of non-native consonant and vowel contrasts have been reported in several previous studies. A well-known example is that native Japanese speakers have difficulties categorising and discriminating the English /r/- /l/ contrast which is absent in Japanese (MacKain et al., 1981; Miyawaki et al., 1975; Sheldon & Strange, 1982; Yamada & Tohkura, 1992). In addition, it was found that native American English (AE) speakers categorised two non-native bilabial stops that are phonologically contrastive in Zulu (voiced plosive /b/ vs. implosive /ɓ/) as nearly equivalent exemplars of the single native bilabial stop /b/ (Best, McRoberts & Goodell, 2001).

Native versus non-native phonetic differences also influence non-native perception according to PAM principles. For example, although both AE and Japanese have a phonological contrast between /r/ and /w/, the phonetic realisation of the contrast differs in the two languages: /r/ is a central dorsal approximant [ɹ] in AE and an alveolar tap /ɾ/ in Japanese, and /w/ is lip-rounded [w] in AE but unrounded [ɯ̟] in Japanese. Japanese listeners gave more /w/ responses than AE listeners in categorisation of an English /w/-/r/ continuum and their categorisation boundary was less steep because the Japanese listeners assimilated the rounded approximant AE /r/ as a poor exemplar of Japanese /w/ rather than as Japanese tapped /r/ (Best & Strange , 1992). Similarly, when French listeners perceived an English /w/-/r/ continuum (/w/-/r/ is also contrastive in French), they categorised /r/ less consistently than AE listeners because /r/ in French is a uvular approximant or

trill and thus differs phonetically from English /r/. Thus, they tended to report the AE /r/ as a /w/, like the Japanese did (Hallé, Best & Levitt 1999). These and other analogous findings suggest that even when there is a similar *phonological* contrast in the listener's native language, *phonetic* differences from the native phonemes also affect perception of the non-native contrast.

5.1.2 Phonological features of Thai, Mandarin, Northern and Southern Vietnamese tones

The four tone languages used for the current study, the stimulus target language Thai (five tones) and three non-native listener languages, Mandarin (four tones), Northern Vietnamese (six tones) and Southern Vietnamese (five tones), differ phonologically in terms of the number and types of tones they contrast. Given that no universal cross-language phonological model of tone systems yet exists, we developed a parsimonious phonological system for our purposes of making phonological predictions for non-native tone assimilations. It includes perceived abstract pitch contours and heights. The contour specifications are level (flat) contour and the dynamic contours rising, falling, and their combinations of falling-rising and rising-falling. The height specifications are high, mid, low. This system minimises dependence on specific, detailed phonetic realisations as actual F0 patterns over time, and is capable of capturing tone distinctions for a wide range of tone languages (perhaps all). Contour (level or dynamic) is specified for each tone in a given language, but height is specified only when it is phonologically contrastive for tones of the same contour, e.g., height distinctions between two falling or two rising or two level tones in the same language. Note that among the four languages selected for the current study, the only height contrast for falling-rising contours occurs in Northern Vietnamese, whereas Mandarin tones (in citation form[12]) contrast only in contour type, with no minimal phonological height contrasts.

[12] The focus of this paper is on perception of monosyllabic tones in citation form. We did not present tones in connected/conversational speech, as this is not typically done in perceptual assimilation studies, and moreover it

Thai, the target stimulus language, has three phonologically level tones contrasting in height (high, mid and low), and two dynamic contour tones that do not contrast in height (rising and falling) (Gandour, 1978). Mandarin has one level and three dynamic contour tones（a rising tone, a falling-rising tone, and a falling tone) in citation form (Chao, 1968).

The tone systems for the two dialects of Vietnamese are more complex. Northern Vietnamese has two phonologically level tones contrasting in height, high level *ngang* and low level *huyền*. It has four phonologically contour tones: rising *sắc*, falling *nặng*, and falling-rising tones that contrast in height: high falling-rising *hỏi*, and low falling-rising *ngã* (Nhàn, 1984). In the process of tonogenesis in Vietnamese (Haudricourt, 1954), in syllables that ended with a glottal stop, that stop evolved into the rising *sắc* and falling *nặng* tones, which differ according to whether the syllable's onset consonant was voiceless versus voiced. Those two simple dynamic tones are phonologically distinct from the dynamically more complex falling-rising tones *hỏi* (high) and *ngã* (low) (determined by the voiceless versus voiced initials), which instead evolved from syllables ending with -s or -h. The reason for specifying *hỏi* as high and *ngã* as low is phonologically motivated (Yip, 2002): in tone reduplication in Northern Vietnamese, if the inputs are rising *sắc* or high falling-rising *hỏi*, the prefixal reduplicant surfaces as high level *ngang*, whereas if the inputs are falling *nặng* or low falling-rising *ngã*, then the prefixal reduplicant surfaces as low level *huyền* (Nhàn, 1984; Yip, 2002).

Southern Vietnamese has five tones in its phonological system. Four of them correspond phonologically to Northern Vietnamese: high level *ngang*, low level *huyền*, rising *sắc*, falling

introduces tone coarticulation and tone sandhi that affect the shapes of the tones. The latter effects are certainly of interest for future research but are beyond the scope of the current study, which already has a complex multifactorial design.

nặng. The fifth tone is reported to reflect a diachronic tone merger of the two falling-rising tones, high *hỏi* and low *ngã* (Brunelle, 2009), implying that the height distinction for falling-rising tones has been lost over time in Southern Vietnamese, which thus retains a height contrast only for level tones.

5.1.3 Phonetic characterisations of Thai, Mandarin, Northern and Southern Vietnamese tones

Several conventions exist for transcribing the phonetic details of lexical tones, including the diacritic convention used by the IPA (The International Phonetic Alphabet, 2015) and Chao numbers (Chao, 1930). As Chao numbers offer more flexibility in characterising dynamic contour tones and have been more widely used in studies of Asian tone languages, we relied on existing Chao transcriptions to provide a first approximation of phonetic characteristics of the tones in our selected languages. In Chao tone transcriptions the number 5 represents the highest pitch in tone production and the number 1 the lowest, and the sequence of Chao numbers indicates the pitch contour of the tone (Chao, 1930). We designated T for Thai, M for Mandarin, NV for Northern Vietnamese and SV for Southern Vietnamese. Thus, Mandarin high level tone is designated in Chao numbers as M55, rising tone as M35, falling-rising tone as M214 and falling tone as M51(Chao, 1968).

The three phonologically level tones in Thai are not, however, all transcribed as being *phonetically* level according to their designated Chao numbers. The three Thai phonologically level tones are phonetically transcribed in Chao numbers as T45 (high level), T33 (mid level), T21 (low level) (Reid et al., 2015). The phonologically rising tone is phonetically transcribed in Chao numbers as T315 and the phonologically falling tone as T241 (Reid et al., 2015).

Northern Vietnamese phonologically level tones are phonetically transcribed in Chao numbers as NV44 high level (*ngang*) and NV22 low level (*huyền*), and the four dynamic contour tones as

NV35 rising (*sắc*), NV21 falling (*nặng*), NV214 high falling-rising (*hỏi*), and NV415 low falling-rising (*ngã*) (Nhàn, 1984)[13]. Four of the five Southern Vietnamese tones are transcribed like their Northern Vietnamese counterparts: SV44 high-level (*ngang*), SV22 low-level (*huyền*), SV35 rising (*sắc*), SV21 falling (*nặng*). The merged falling-rising tone (merger of NV *hỏi* and *ngã*) is transcribed as SV214 (Brunelle, 2009).

Note the non-level Chao phonetic transcriptions for two of the Thai phonologically level tones and the falling-rising and rising-falling Chao transcriptions for the phonologically rising and falling Thai tones, respectively. These cases highlight that there are discrepancies between the phonological characterisation of tones and the Chao transcriptions of their phonetic realisations. In addition, even the Chao transcriptions may also mismatch acoustic measures of the tones' actual F0 contours. These observations highlight the difficulties of comparing the phonetic characteristics of tones across languages using Chao transcriptions alone. Therefore, we also conducted acoustic analyses of recorded tokens in each language to determine their detailed acoustic properties.

5.1.4 Acoustic properties of lexical tones

Lexical tones are realised primarily by F0 variations[14] and are usually characterised acoustically as either F0 contours or discrete features derived from F0 contours. We followed that standard

[13] The designation of high versus low features in this contrast is based on their phonological behaviour in reduplication as stated in the phonological description section, and there is a discrepancy between the phonological features and Chao numbers in this case.

[14] Other acoustic properties, such as voice quality, have been claimed to be associated with some tones in some languages, such as Vietnamese. For example, Pham (2004, 2003) includes phonation type as part of the phonological specification for some northern Vietnamese tones. However, these other features are not proposed to be required for tone languages generally nor specified for most tone languages that have been described. As our aim is to present a universal model of tone features, we focus on F0 properties, which are central to descriptions of every known tone language, in the present paper.

approach. We recorded production of two consonant-vowel syllables (/ma/ and /mi/) by four female native speakers of each language (Mandarin, M_{age} = 27.0 years; Northern Vietnamese, M_{age} = 22.5 years; and Southern Vietnamese, M_{age} = 20.5 years; Thai, M_{age} = 30.3 years) with each of their native tones. All target syllables are meaningful (free or bound) morphemes in the respective languages. There were 64 tokens (2 syllables /ma/ and /mi/ × 4 tones × 8 repetitions) for each Mandarin informant and 96 tokens (2 syllables /ma/ and /mi/ × 6 tones × 8 repetitions) for North and South Vietnamese informants. Mandarin and Vietnamese productions were recorded using a Zoom H4n digital speech recorder with a sampling rate of 44.1 kHz 16-bit stereo format at a quiet testing booth at The MARCS Institute, Western Sydney University.

Mandarin items were elicited via Pinyin; Vietnamese and Thai items were elicited via the orthography of their languages in random order. Southern Vietnamese speakers were asked to read all six Vietnamese tones but were specifically instructed to read them in their southern accent, which allowed us to confirm whether the two falling-rising tones in Southern Vietnamese are indeed acoustically merged by our participant cohort. This merger was observed in our acoustic analyses of their tone productions (see Appendix A, Figure A.1). In Northern Vietnamese, the acoustic differences between these two tones are significant $F(5.11) = 18.66$, $p < .001$ but in Southern Vietnamese, the difference is not significant, $F(2.00) = 0.16$, $p = .86$.

The Thai syllables were recorded for a separate study (Burnham et al., 2009), and were used here with permission from the authors. These materials were recorded in citation form in a sound-treated booth at The MARCS Institute, Western Sydney University, using a Lavalier AKG C417 PP microphone at the sampling rate of 48 kHz and 16 bit resolution. Many repetitions were produced by the speakers, but only natural-sounding high-quality exemplars of each token were selected. Five tokens per type were selected from two participants, four from the third participant,

and six from the fourth participant for the current acoustic analysis. Thus, there were 20 tokens for each of the ten Thai target items (/ma:/ and /mi:/ × 5 tones each) for a total of 200 Thai tokens.

All boundaries of the recorded syllables were automatically marked and manually checked. The Praat script *ProsodyPro* (Xu, 2013) was used to provided syllable durations and 10-equidistant points of F0 values (in Hz). Raw F0 (in Hz) was normalised using the Lobanov (1971) method, which is commonly used in studies of, e.g., vowel acoustics across multiple speakers, which requires normalisation of frequency ranges and values across speakers. Lobanov-normalised F0 values reflect variation from the F0 mean of the speaker[15]. The most stable part of the normalised tone (from 10% to 90% of the syllable) was used to calculate all F0-related measures. Figure 5.1 shows the Lobanov- and time-normalised mean F0 contours of lexical tones in Thai, Mandarin, Northern and Southern Vietnamese. For more details on discrete measures, such as syllable duration and F0-related measures, see Appendix A (Table A.1).

[15] Calculating semitones instead, as suggested by one reviewer, would not have yielded the cross-speaker F0 normalisation needed for making cross-language tone system comparisons.

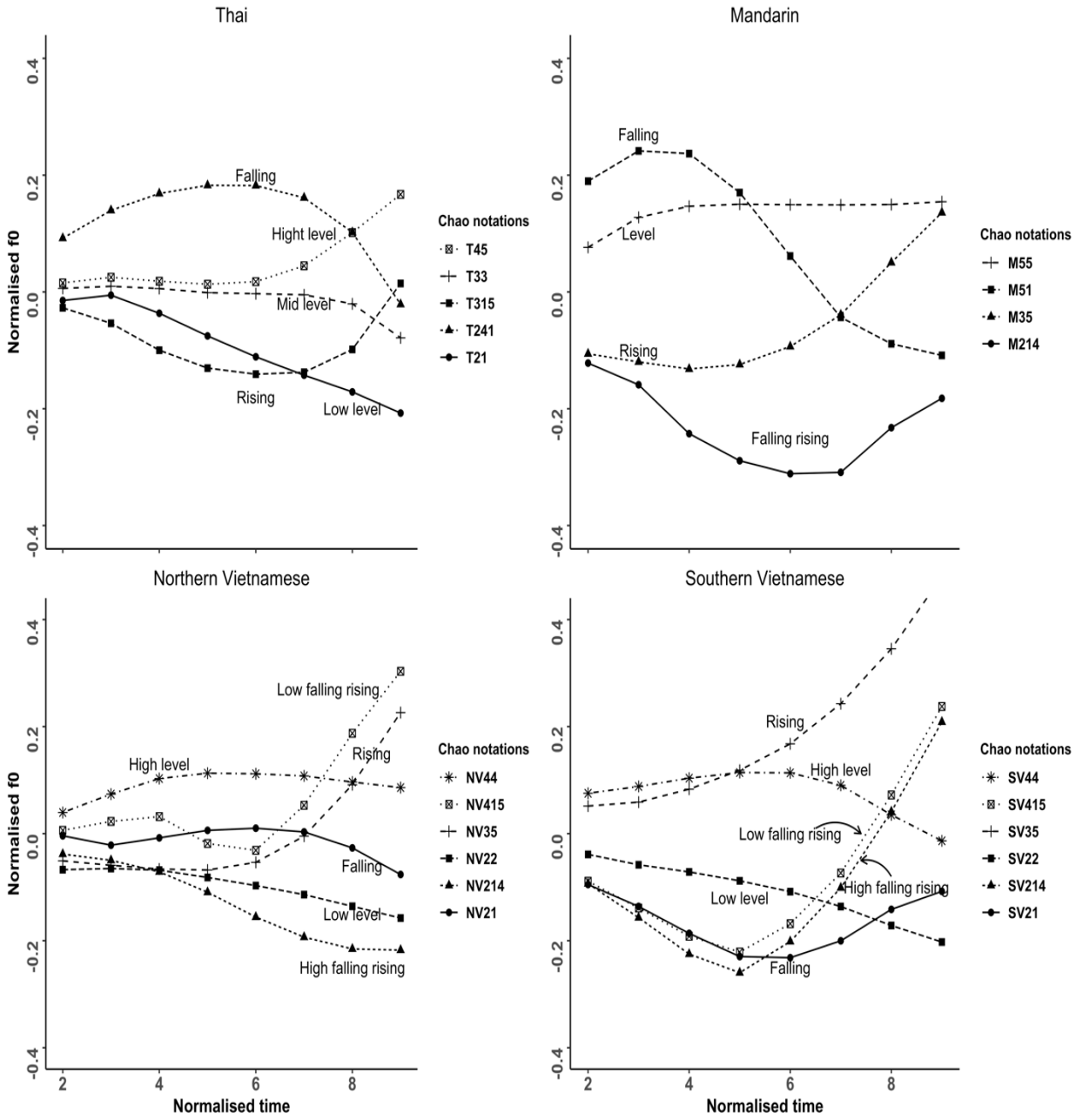


Figure 5.1 Time-normalised and Lobanov-normalised mean F0 contours of Thai (upper left), Mandarin (upper right), Northern Vietnamese (lower left) and Southern Vietnamese tones (lower right). Averaged across all tokens per tone type in each language. Thai: 20 tokens × 2 syllables (/ma:/ or /mi:/) × 5 tones; Mandarin: 32 tokens × 2 syllables (/ma/ or /mi/) × 4 tones; Northern Vietnamese/Southern Vietnamese: 32 tokens × 2 syllables (/ma/ or /mi/) × 6 tones. Phonological type labels are provided in the graphs for each contour.

5.1.5 PAM predictions

The basic PAM principle (Best, 1995) is that naïve listeners will assimilate non-native phones into their native phonological categories but with differences in goodness of fit due to the magnitude of phonetic discrepancies from the native phonemes. A non-native phone can be a good to moderate to poor fit (phonetic aspect, which is gradient) to a native phonological category (phonological aspect, which is categorical). Following these PAM principles, we posit that the extent of native phonological influence is reflected in the assimilation types. Phonological influence is generally strong for Categorised assimilations and is weaker for UnCategorised assimilations with variations for subtypes. In UnCategorised$_{focalised}$ assimilations, the non-native phone is still assimilated as primarily similar to a single native category but choices of that native phoneme fall below the defined categorisation threshold and thus the phonological influence is moderate. In UnCategorised$_{clustered}$ assimilation, the non-native phone is assimilated to a small set of native categories (below threshold but above chance level) and thus they each have a weak influence on assimilation. The general native phonological influence is therefore weak. In the UnCategorised$_{dispersed}$ assimilation, a non-native phone category is spread across many L1 categories (all below chance level). The native phonological influence is virtually non-existent in this case.

Within those phonological constraints, listeners will nevertheless display some residual sensitivity to within-category *phonetic* variations from their native categories commensurate with the magnitude of phonetic discrepancy from the native tone(s) it is assimilated to. Residual native phonetic sensitivity is determined separately based on percent choice and goodness ratings of the chosen categories. We divided percent choice of the native tones above chance into three ranges: Low, Medium and High. These ranges respectively reflect strong, moderate, and

weak residual phonetic effects. The percent choice ranges necessarily differ for the three groups. For Mandarin listeners, Low spanned 25% (chance level) to 49% of choices, Medium spanned 50-75%, and High spanned 76-100%. For Northern Vietnamese listeners, Low covered 17% (chance level) to 44%, Medium spanned 45-73%, and High spanned 74-100%. For Southern Vietnamese listeners, Low spanned 20% (chance level) to 46%, Medium spanned 47-74%, and High spanned 75-100%. For the category-goodness ratings, we also divided the scale in to three ranges, which apply to all listener groups: Low = 1-2.9, Medium = 3-4.9, and High = 5-7. These ranges reflect strong, moderate and weak residual phonetic effects, respectively.

High percent choice and/or high ratings reflect very strong perceived similarity to the corresponding native category and very weak residual phonetic sensitivity to within-category *phonetic* variations from their native categories. Low percent choice and/or low ratings, on the other hand, reflect very weak perceived similarity to the native tone and very strong residual phonetic effects. Percent choices reflect residual phonetic sensitivity in the categorisation process, which focuses on identifying how non-native phones may correspond to native phonological categories. Goodness of fit ratings instead reflect residual phonetic sensitivity in the separate goodness rating process that focuses on evaluation of within-category phonetic differences between the non-native phone and native realisations of the category(s) to which it is assimilated. Thus, it is possible to observe discrepancy between percent choice and ratings within Categorised and UnCategorised assimilation, e.g., high-range percent choice with medium-range ratings or vice versa.

High and medium ranges of both percent choice and goodness ratings are predicted to occur for Categorised assimilations because they inherently entail perceived strong to moderately strong global phonetic similarities to the “most similar” native categories. Medium and low ranges of

both percent choice and goodness ratings are predicted to occur for UnCategorised assimilations because the assimilated responses inherently presuppose moderately weak to much weaker perceived global phonetic similarities.

The only level of percent choice and ratings that overlaps between Categorised and UnCategorised$_{focalised}$ (possibly also UnCategorised$_{clustered}$) is "Medium", which would refer to Categorised cases that do meet the statistical criterion but with moderate levels of percent choice, as well as UnCategorised$_{focalised/clustered}$ cases that fail the statistical criterion for Categorisation but still have moderate levels of percent choice for one or more native categories.

We extended these PAM principles to assimilation of non-native lexical tones by naive tone language listeners to make predictions about non-native Thai tone assimilation by listeners of three tone languages/dialects. It is important to reiterate our assumption that naïve listeners lack phonological knowledge of the target language and thus are phonologically constrained only by their native tone system. Thus, we considered both perceived abstract pitch contours and heights and phonetic characteristics for the native tones but only the phonetic characteristics for the non-native tones. For making predictions for phonetic effects, I combined acoustic measures (such as syllable duration and F0-related measures, see Appendix A Table A.1) and F0 properties as in Figure 5.1.

T45 is phonetically rising in Thai and therefore should be Categorised as M35 because the general phonetic form of T45 fits abstractly the phonological feature specifications of M35 rising tone. The percent choice and category-goodness rating of M35 should be in the high range, reflecting a low residual sensitivity to non-native phonetic variations from the corresponding native category.

As noted earlier, Mandarin has only one level tone with no contrast in height. Thus, the Thai level tone T33 is expected to be Categorised as M55. The percent choice should be high which reflects

low residual phonetic sensitivity in phonological categorisation to the only level tone in the native system, whereas the goodness rating should be medium as the phonetic evaluation of T33 in the rating task should reveal residual sensitivity to the moderate within-category phonetic deviation of T33 from M55 in terms of height.

T21 is phonetically low with a falling contour, but Mandarin does not have a phonologically low falling category. However, M214 is often realised phonetically as a low falling (21) allotone in non-final position although it acquires a final rise in the citation form (Yip, 2002). Thus, T21 should be Categorised as M214. However, the percent choice and rating for M214 as a response category will be medium given the notable departure from the native tone's final rise in citation form (as in our stimulus materials). M51 should also be selected as it is the actual falling tone in Mandarin. But the phonetic differences between M51 and T21 will lower the percent choice and category-goodness ratings to the low range.

T315 is phonetically falling-rising. Both M35 and M214 are phonetically similar to T315, in different ways, but M35 is more similar in terms of initial and final height. Thus, it should be Categorised as M35 with medium percent choice and ratings, but M214 should be selected less often, with a low percent choice and category-goodness rating.

T241 is phonetically rising-falling and because Mandarin does not have such a phonological category, it is expected to be UnCategorised. Due to its greater phonetic similarities (height and contour) to M51 and somewhat to M55 (height) than to M35 or M214, as shown in Figure 5.1, the percent choice of each of the former two categories should be higher than those for the latter two, resulting in UnCategorised$_{clustered}$ assimilation. The percent choice and category-goodness rating of M51 should be medium, given the moderate phonetic similarity, and M55 choices and ratings should be in the low range.

For Northern Vietnamese listeners, T45 is phonetically rising and is similar to native rising tone NV35 but has a less steep rise with higher initial F0 and lower final F0, so it should be categorised to NV35 but with medium percent choice and goodness ratings. NV415 is also similar to T45 in its initial and final F0 but NV415 has a dip in the middle of the tone contour whereas T45 lacks such a medial dip. This phonetic difference from T45 should lead to some selection of NV415 but with low percent choice and low category-goodness ratings.

For Southern Vietnamese listeners, T45 should also be assimilated to SV35. However, SV214 (the merged falling-rising tone) and SV21 are also phonetically similar to T45 in terms of F0 contour, though with a lower overall F0 height (see Figure 5.1). Given that SV214 and SV21 contours deviate from that of T45, and SV lacks a height contrast in contour tones, it is predicted that T45 will be UnCategorised and split among SV35, SV21 and SV214, thus $\text{UnCategorised}_{\text{clustered}}$/$\text{UnCategorised}_{\text{dispersed}}$. In this case, the percent choice of each of these SV tones should be in the low range, but the ratings should be in the medium range due to somewhat greater phonetic similarity.

Northern Vietnamese and Southern Vietnamese each have two level tones that contrast phonologically in height. T33 is phonetically lower than NV/SV44 but is only slightly above the height of NV/SV22, although it also has a somewhat different final contour. Thus, we predict the T33 would be Categorised as NV/SV22 with medium percent choice and category-goodness ratings. NV/SV44 should also be selected but the phonetic differences will lead to low percent choice and category-goodness ratings.

For low level T21, Northern Vietnamese and Southern Vietnamese both have a low level tone, NV/SV22, that is phonetically similar to T21. Thus, T21 should be Categorised as NV/SV22 with high percent choice and category-goodness ratings. The Vietnamese falling tones NV/SV21 are

also phonetically somewhat similar to T21 but each of their contours differs phonetically from T21 as shown in Figure 5.1, so they should be selected with a low percent choice and category-goodness rating.

T315 is phonetically falling rising and Northern Vietnamese has two phonologically falling rising tones, NV214 and NV415 but only NV415 has a final rise, as shown in Figure 5.1. NV35 also has a similar final rise but without a dip in the middle of the tone. But NV415 is higher than NV35 overall and is steeper in the final rise of the contour and thus more deviant from T315 than NV35. Thus, T315 should be Categorised as NV35 with medium percent choice and goodness ratings, while NV415 will also be selected but with a low percent choice and category-goodness ratings.

On the other hand, Southern Vietnamese has only one falling rising tone, which reaches a low but not extremely low F0 in the middle. Thus, T315 is predicted to be Categorised as SV214 with medium percent choice and goodness ratings. SV35, with similar initial and final F0 but no dipping part in the middle of the contour, is less similar to T315 than SV214. It will be selected with low percent choice and low goodness ratings.

T241 is phonetically rising falling and both Northern Vietnamese and Southern Vietnamese lack rising falling tones in their tone inventories. However, NV/SV44 are phonetically similar to T241 in F0 height and throughout much of their contours (in Figure 5.1). Thus, T241 is predicted to be Categorised as NV/SV44; however, percent choice and category-goodness ratings should be medium.

Native phonological and phonetic influences will also be reflected in response times in the categorisation task. When a non-native tone is categorised to the native tone system, the incoming stimulus token is assumed to be stored in working memory (Baddeley, 2010; Baddeley & Hitch, 1974). If a non-native tone is phonologically Categorised and phonetically an ideal exemplar of

the native tone, i.e., this tone falls squarely within that native phonological category, then the native category should be highly activated and receive a high percent choice. No other native phonological categories have been activated and thus do not compete with it in the categorisation process. Thus, response times in this case should be short.

However, if a non-native tone is phonologically Categorised yet phonetically deviant from the native category, response times should be longer. That is, the processing cost should increase due to the phonetic discrepancy from the native tone, which would delay the decision (response time). In addition, if a non-native tone resembles two or more native categories, i.e., is assimilated as UnCategorised, then those multiple categories will be activated, each more weakly than in a categorised non-native phone. This should also impose an even greater processing cost due to the competition between the alternatives, which will delay the final categorisation decision. At the group level these added processing costs should yield selection of different (partially) activated native categories from trial to trial, as well as yielding longer average response times due to increased category uncertainty.

Although the effects of talker variability on response time in speech processing have been examined (Antoniou & Wong, 2015), response time differences between Categorised and UnCategorised assimilation types have not previously been investigated. Within UnCategorised assimilation, as noted earlier, we further distinguish among three types of responses (*focalised, clustered* and *dispersed* responses), following Faris and colleagues (2018). We predict that *focalised* responses, in which there is one major native response category, should yield relatively shorter decision times than *clustered* responses in which there are more than one native response categories activated and thus more competition. In a similar vein, *dispersed* responses should result in the longest decision times as most/all native categories are activated to a low and roughly

equivalent level, that is, none of them are a good match for the non-native phone. The high phonological and phonetic uncertainty will maximise decision time and variability of categorisation choices.

5.2 Method

5.2.1 Participants

Twelve[16] native speakers of Mandarin (M_{age} = 29.6 years, *SD* = 6.1; 10 females), Northern Vietnamese (M_{age} = 20.4 years, *SD* = 2.5; 6 females), Southern Vietnamese (M_{age} = 23.7 years, *SD* = 8.6; 10 females) participated in the experiment. Mandarin participants were tested at a university in Nanjing, China, except for three who were on a six-month academic visiting program in Sydney when they participated in the experiment. The Mandarin-speaking participants were all born and raised in various dialect regions in China (e.g., Fujian, Jiangsu, Jiangxi, Liaoning). All of their education had been conducted in Mandarin, spanning from early childhood through university, and they continued to use Mandarin on a daily basis. The Vietnamese participants were tested at universities in Sydney, Australia, where they were enrolled. Vietnamese participants were born and raised in the Northern regions (around Hanoi) or the Southern regions (around Ho Chi Minh city) of Vietnam and came to Australia mostly to study or migrate (Northern participants' $M_{age\ of\ arrival}$ = 19.7 years, *SD* = 2.3; Southern participants' $M_{age\ of\ arrival}$ = 23.2 years, *SD* = 8.8). According to a background questionnaire completed before the test, all had less than two years of formal

[16] We originally tested 13 participants in each group, but each group had one heritage speaker, so we conducted the analyses both with all speakers (*n* = 13 per group), and with the heritage speakers removed (*n* = 12 per group) for comparison. The only noteworthy difference is that the choice of M214 for T21 was lowered from 62.2% to 59.1% and that of M51 was raised from 25.7% to 27.9%, and consequently the difference between the two became marginal (*p* = 0.06). We reported results with a stricter control of listener backgrounds (n = 12 per group) but still considered T21 to be Categorised as M214.

musical training, which is important because more extensive musical training influences tone perception (Gottfried et al., 2004). All participants self-reported to have normal hearing, no known speech-language difficulties, and none had experience with Thai[17]. The experiments were approved by the Western Sydney University Human Research Ethics Committee (HREC12560). All participants signed a consent form prior to testing and were compensated for their time (AU$15 or the equivalent 60 RMB for participants in Nanjing, for 45 minutes).

5.2.2 Stimulus materials

We used a subset of the same monosyllabic Thai stimuli as in the acoustic analyses for the perceptual stimuli, with permission from the original authors (Burnham et al., 2009). Two syllables (/maː/, /miː/, with long Thai vowels) were chosen for target stimuli because they form (free/bound) morphemes for each native tone in Thai, Mandarin and Vietnamese, as noted earlier (*1.4 Acoustic properties of lexical tones*). We gave participants explicit instructions to assimilate the tones into their native tone categories (not to identify them as native words). This should minimise the effects of different morphological status. For the perceptual stimuli, we selected two tokens of each Thai target item by two native female speakers that had been judged by a third native Thai listener as the most natural sounding and correct.

5.2.3 Procedure

Participants were tested individually in a quiet room (e.g., sound-attenuated testing booth at Western Sydney University; library study booths at Macquarie University, University of Technology Sydney, and Nanjing University). Stimuli were presented on a Dell Latitude 7280

[17] They had begun learning English as school children in their home countries. Note that experience with English as a second language would not modulate their native tone system phonological knowledge, as English is not a tone language.

laptop running E-Prime Professional 2. Auditory stimuli were presented via Sennheiser HD 280 Pro headphones at 72 dB SPL.

Before the test session, participants completed 20 practice trials to familiarise them with the task and decisions they were asked to make on each trial. The stimuli in the practice trials were not used in the test session. No feedback was given in either the practice trials or test session.

The categorisation task had 120 trials (2 speakers × 5 tones × 2 tokens × 2 syllables × 3 repetitions). On each trial, the stimulus token was presented, and listeners made a forced-choice categorisation judgment to their native tones via a key press. Mandarin participants chose from four Pinyin options on stickers on keyboard keys, and Vietnamese speakers chose from six Vietnamese transcriptions on stickers on keyboard keys. The stickers were placed on the keys in the same line on the keyboard, i.e., “f”, “g”, “h”, “j” for Mandarin participants, “f”, “g”, “h”, “j”, “k” for Southern Vietnamese participants, and “d”, “f”, “g”, “h”, “j”, “k”. None of the participants reported any problems using these keys which were handy when participants were familiar with their positions. In addition, these keys are quite close to each other, which minimises possible effects on reaction time. Participants were asked to respond as quickly as possible within a 3s response period. Responses beyond 3s were not used for analysis; missing data due to time-outs account for only ~1% of the responses. Immediately following their categorisation decision, the same stimulus was played again in the same trial, and participants were asked to rate how well the tone fitted into the native category they had chosen, on a goodness rating scale (1=Poor and 7 = Perfect). The ratings were used to index perceptual sensitivity to phonetic differences between the non-native token and the chosen native phonological category. Response times were also recorded for the categorisation responses.

5.3 Results

5.3.1 Categorisation criteria

Although in many published studies Categorised assimilation has been defined as the selection of one native category above an arbitrary pre-set threshold (Tyler et al., 2014) for a given non-native phone, we followed the tone categorisation criteria used in So and Best (2014) as being more appropriate for categorisation of lexical tones by listeners whose native languages differ in number of tones: first, a given native tone must be selected significantly more than chance level; second, that single native tone category must be chosen significantly more often than any other native categories. This method is more sensitive to variations among different native tone systems than a rigid threshold is, as it considers the number of tone categories in a particular listener language.

For UnCategorised assimilations, we followed the subtypes established by the two previous studies of Faris and colleagues (2016, 2018), but modified their operational criteria to be based on our statistical approach to categorisation (Faris et al. had defined them using the fixed threshold approach) as follows: (1) for a focalised response, one non-native phone was considered as primarily similar to a single L1 category（above chance level）but choices of that native phoneme were not significantly higher than choices of other native categories; (2) for a *clustered* response, the uncategorised non-native phone was assimilated to two or more L1 categories above chance level, but they were not significantly different from each other; or (3) for a *dispersed* response, the choice of native phone category was made randomly across many L1 categories, all below chance level and not significantly different from each other.

Note that choices of SV214 (*hỏi*) and SV415 (*ngã*) (orthographically different) by the Southern Vietnamese group were combined for analyses of categorisations, ratings and reaction times,

because they are phonologically merged rather than contrastive in Southern Vietnamese (see the evidence of merger in 1.4 Acoustic properties of lexical tones).

5.3.2 Percent choice and rating scores in perceptual assimilation

The mean percent choices and category-goodness ratings for each native tone response category are shown in Figure 5.2 (for the full set of percent choices including those below 1% and the category-goodness ratings of below-chance choices, see Appendix A, Table A.2).

To determine whether an assimilation was Categorised or UnCategorised, we assessed whether categorisations of each Thai tone were significantly above chance level, which is 25% (1/4) for Mandarin, 16.7% (1/6) for Northern Vietnamese, and 20% (1/5) for Southern Vietnamese speakers, via a series of t-tests (see Appendix A, Table A.3, for statistical details).

Then, for each listener group, linear mixed effects models were built for each of the five Thai target tones to determine whether the listeners assigned different native tone category labels to the Thai targets (details in Appendix A, Table A.4). Participants were specified as a random factor. To calculate the *p*-values for the fixed effects, we used the Kenward-Roger approximation to the degrees of freedom, as recommend by Halekoh & Hojsgaard (2014), and the *Anova* function from the *car* package in R, with test specified as "F".

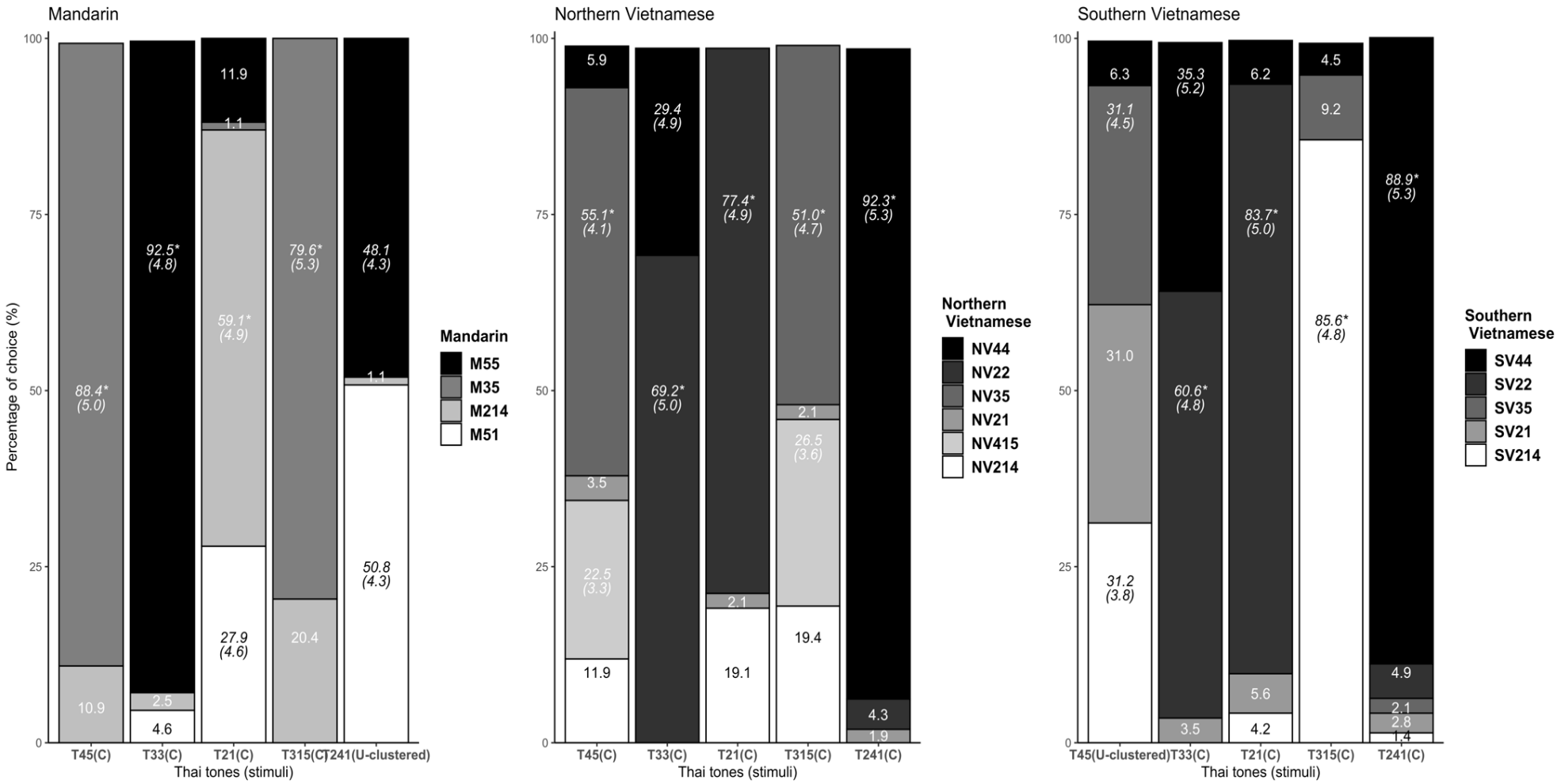


Figure 5.2 Mean percent choices and goodness ratings (in parentheses) for each Thai tone by Mandarin, Northern Vietnamese and Southern Vietnamese listeners. Note: Categories in italics are choices that were significantly above chance, which was 25% for Mandarin, 16.7% for Northern Vietnamese, 20% for Southern Vietnamese; "*" = Categorised tone. Assimilations: C = Categorised, U = UnCategorised. Ratings are shown for above chance level responses: 1 = poor, 7 = perfect. Native response categories less than 1% are not shown here but can be found in Appendix A, Table A.2.

Next, we ran multiple comparisons with the R-package *lsmeans* to determine which native category was selected more than other native categories that were above chance level (see Appendix A, Table A.5, for statistical details). A Thai tone was considered Categorised only when one native category was selected significantly more often than all other native categories. We also ran t-tests on category goodness ratings differences for cases where two non-native tones were assimilated into the same single native category. This allowed us to distinguish Single-Category assimilation (no significant difference in category goodness ratings) from Category-Goodness assimilation (significant rating difference between the two non-native tones), following PAM principles.

For Mandarin listeners, four of the five Thai tones were Categorised. Both T45 and T315 were Categorised as M35, but their category goodness ratings were not significantly different, thus meeting the Single-Category assimilation criterion. T33 was Categorised as M55. In all three of these cases only one native category was selected significantly above chance level and significantly more often than all other native choices. The fourth case, T21, was Categorised as M214, which was chosen marginally significantly more often than the other three Mandarin choices, although M51 was also chosen above chance. T241 was instead an $\text{UnCategorised}_{\text{clustered}}$ assimilation because it was split evenly between M55 and M51, and both were chosen above chance level but not significantly different from each other.

For Northern Vietnamese listeners, all Thai tones were Categorised as native tones. T45 and T315 were both Categorised as NV35, and their category goodness ratings were not significantly different, meeting the Single-Category criterion. T33 and T21 were Categorised as NV22, and

their category goodness ratings were not significantly different, again a Single-Category assimilation. T241 was Categorised as NV44.

Southern Vietnamese listeners Categorised both T21 and T33 as their native SV22, and their category goodness ratings were not significantly different, thus this pair forms a Single-Category assimilation. They also Categorised T315 as their SV214 and T241 as their SV44. However, T45 was an $\text{UnCategorised}_{\text{clustered}}$ assimilation, as SV35 and SV214 were selected significantly above chance but not significantly different from each other.

5.3.3 Categorisation response times for the different assimilation types

The mean response times (ms) to each Thai tone for each native choice are presented in Appendix A, Table A.2. For formal statistical modelling, response time data were restricted to the categorised choices for the Categorised assimilations, and to above-chance choices for the $\text{UnCategorised}_{\text{clustered}}$ assimilation,

We predicted that the response time of Categorised assimilations should be shorter than that of UnCategorised assimilations. This is because as a non-native tone is assimilated across more native categories, more native categories will be activated, which should induce greater processing cost due to the comparison among the alternatives, and this will delay the final categorisation decision.

We regrouped response time data across the three listener groups according to assimilation types, and ran a mixed-effects model with response time as the dependent variable and assimilation types (Categorised, UnCategorised) as the fixed factor. Participants were specified as random intercepts. A significant difference was found between Categorised and UnCategorised assimilations in terms of response time, $F(1, 146) = 14.96$, $p < .001$, suggesting that Categorised assimilations ($M = 764$

ms, *SD* = 339 ms) were significantly faster than UnCategorised assimilations (*M* = 928 ms, *SD* = 348 ms) as we predicted.

5.3.4 Residual native phonetic sensitivity

Next, we describe the range of percent choice and goodness ratings for the assimilations (see Table 5.1 for a summary in relation to our predictions).

First, we successfully predicted Mandarin listeners' Categorised assimilation of T45 as M35 and with accurate predictions of percent choice and rating in the high ranges. T33, as predicted, was Categorised as M55 (the only level tone in Mandarin, not contrastive in height) by Mandarin listeners with a high percent choice, indicating a strong native phonological influence and reduced residual phonetic sensitivity in the categorisation process. However, in the rating task, Mandarin listeners did show sensitivity to detecting phonetic differences of T33 from their native level tone M55 as indicated by a medium level category-goodness rating.

In addition, as predicted, T21 was Categorised as M214 with medium range of percent choice and ratings. The assimilation of T315 to M35 was also successfully predicted but the percent choice and ratings were higher than expected, again indicating unexpectedly reduced residual sensitivity to differences from the native tone in both categorisation and rating processes, i.e., overridden by a strong native phonological influence.

T241 was predicted to be UnCategorised and split between M51 and M55. The M51 selections were expected to display medium range scores in percent choice and goodness ratings whereas M55 choices were expected to show low range scores in percent choice and goodness ratings. The Mandarin listeners in present study did split their responses to T241 between M51 and M55, yielding an $\text{UnCategorised}_{\text{clustered}}$ assimilation, but percent choice and ratings were roughly equal between the two. This finding suggests that Mandarin listeners may have perceived the initial

contour (24) in T241 as level rather than as rising. Previous reports have also shown some variations in assimilation of T241 by Mandarin listeners. Notably, previous observations encompass both of the assimilations we observed here: whereas both experienced (Thai learners) and inexperienced (Thai-naive) Mandarin listeners in Wu et al. (2014) categorised T241 as M51, Mandarin participants in the study reported by Reid and colleagues (2015) categorised T241 as M55.

For the Northern Vietnamese listener group, our predictions based on PAM principles were also generally upheld. Northern Vietnamese listeners Categorised T45 as NV35, as predicted. The phonetic differences between T45 and NV35, as expected, lowered the percent choice and ratings to medium range.

The prediction that both T33 and T21 would be Categorised to NV22 was supported. The percent choice of NV22 for T33 was smaller than that of T21, indicating larger residual sensitivity to T33 than T21 phonetic differences from NV22 in the categorisation process. The category-goodness ratings for both T33 and T21 were comparable, both reflecting relatively medium sensitivity to phonetic variations of the native categories from NV22 in the rating process.

We successfully predicted that T315 should be Categorised as NV35, with lower choices of NV415. Both the percent choice and category-goodness rating of NV35 were in the medium range, as we had predicted. NV415 was also selected as a native response category, but with low percentage of choice and medium category-goodness ratings. The rating was higher for NV35 than NV415, indicating that the residual sensitivity to phonetic differences from T315 was greater for NV415 than NV35.

T241 was Categorised as NV44 as predicted, but with higher than predicted percent choice and ratings. This reflects strong native phonological constraints and relatedly less phonetic sensitivity

to the deviation of the non-native tone from the native one in both categorisation and rating processes.

Also consistent with our PAM-derived predictions, Southern Vietnamese listeners behaved quite differently from Northern Vietnamese listeners when assimilating the Thai rising contour tones T45 and T315. T45 was split by Southern Vietnamese listeners among SV35, SV214 and SV21, as expected. The percent choices were in the low range as predicted and the goodness ratings for all three native response categories were in the medium range, also as predicted. As we said, both SV21 and SV214 are phonetically falling-rising, and SV214 and SV415 have been subsumed into a tone merger (unlike the contrast maintained between Northern Vietnamese NV214-NV415: Brunelle, 2009). This difference between the two regional varieties of Vietnamese can account for the different assimilation patterns of T45 by Northern Vietnamese and Southern Vietnamese listeners.

On the other hand, as predicted, Southern Vietnamese participants Categorised T33 as SV22 in a similar way as Northern Vietnamese listeners but with lower percent choice and goodness ratings, suggesting stronger residual phonetic sensitivity to within category phonetic deviations of T33 from SV22 than Northern Vietnamese listeners. T21 was Categorised as SV22 by Southern Vietnamese listeners with similar percent choice and goodness ratings to those of Northern Vietnamese listeners.

T315 was Categorised into SV214 by Southern Vietnamese listeners with percent choice in the high range, higher than predicted (medium range), suggesting reduced residual sensitivity to phonetic differences between T315 and SV214 in the categorisation process, relative to the goodness rating process.

Table 5.1 Summary of phonological and phonetic predictions and experiment results. Assimilations: C = Categorised, U = UnCategorised. High (H), Medium (M) and Low (L) ranges for Categorised assimilations: for Mandarin speakers, L = 25%-49%, M = 50%-75%, H = 76%-100%; for Northern Vietnamese listeners, L = 17%-44%, M = 45%-73%, H = 74%-100%; and for Southern Vietnamese listeners, L = 20%-46%, M = 47%-74%, and H = 75%-100%. For UnCategorised assimilation, the range is from Medium to Low and to Below chance. For ratings, Low = 1-2.9, Medium = 3-4.9, High = 5-7. Results higher than predictions are in boldface, indicating lower than expected within category residual phonetic sensitivity.

	Mandarin			Northern Vietnamese			Southern Vietnamese		
Thai	Predictions		Results	Predictions		Results	Predictions		Results
	Phonological	Phonetic		Phonological	Phonetic		Phonological	Phonetic	
		(% ratings)	(% ratings)		(% ratings)	(% ratings)		(% ratings)	(% ratings)
T45	C as M35	H H	C as M35 (88.4 5.0)	C as NV35	M M	C as NV35 (55.1 4.1)	$UC_{clustered/dispersed}$ SV35 SV214 SV21	 L M L M L M	$UC_{clustered}$ (31.1 4.5) (31.2 3.8) (31.0 4.7)
T33	C as M55	H M	C as M55 (92.5 4.8)	C as NV22	M M	C as NV22 (69.2 **5.0**)	C as SV22	M M	C as SV22 (60.6 4.8)
T21	C as M214	M M	C as M214 (59.1 4.9)	C as NV22	H H	C as NV22 (77.4 4.9)	C as SV22	H H	C as SV22 (83.7 5.0)
T315	C as M35	M M	C as M35 (**79.6 5.3**)	C as NV35	M M	C as NV35 (51.0 4.7)	C as SV214	M M	C as SV214 (**85.6** 4.8)
T241	$UC_{-clustered}$ M51 M55	 M M L L	$UC_{-clustered}$ (50.8 4.3) (48.1 **4.3**)	C as NV44	M M	C as NV44 (**92.3 5.3**)	C as SV44	M M	C as SV44 (**88.9 5.3**)

5.4 Discussion

The PAM-based predictions were generally upheld at both phonological and phonetic levels. Specifically, non-native listeners showed strong evidence of native phonological influences in categorisations, as predicted, as well as residual phonetic sensitivity to non-native tones that reflect native phonetic influences, that were taken into consideration in our extrapolation of PAM principles to perception of non-native tones by native listeners of other tone languages.

We argued earlier that native phonological factors are indicated by Categorised (strong native phonological influences) or UnCategorised assimilations (weak native phonological influences) to the listener's native tone system. Phonetic factors, on the other hand, are reflected in the relative percent choice in the categorisation process, and/or the goodness ratings in the rating process for a native tone category. We considered both the type of assimilation (Categorised versus UnCategorised) and the relative percent choice and goodness ratings to disentangle phonological versus phonetic contributions from the native language on non-native tone categorisation.

In the experiment, we found only two UnCategorised$_{clustered}$ assimilations, reflecting weak native phonological influences whereas all other assimilations were Categorised which suggests that native language phonological constraints are strong in cross-language tone categorisation. Another piece of evidence for native phonological influences is that in some Categorised assimilation cases, the percent choice and/or category-goodness ratings were somewhat higher than predicted (seen in T33→M55; T315→M35 and SV214; T241→NV/SV44). In these cases, it appears that strong phonological influences in Categorised assimilation also reduced residual sensitivity to phonetic deviations of the non-native tones from the native tones. This finding is consistent with the PAM

principle that native phonological influences may hinder sensitivity to within-category phonetic differences. Both the rarity of Uncategorized assimilation and reduced residual phonetic sensitivity are also compatible with claims by NLM that tokens near a native prototype (most phonetically similar to it) are the most greatly perceptually attracted to the prototype with reduced sensitivity to phonetic differences and thus less likely to be UnCategorised, but the farther away the tokens are from the prototype (the more deviant they are from the prototype) the more their phonetic differences are detected.

Strong evidence for phonetic effects, i.e., strong residual sensitivity to phonetic variation from native categories, are indicated by low-range percent choice and/or goodness ratings (seen in T241→M55, T45→SV35/SV214/SV21). More interestingly, in some cases, we found a discrepancy between percent choice and goodness ratings. For example, Mandarin has only one level tone M55 and thus does not contrast level tone height phonologically as Thai does. Mandarin listeners consistently Categorised mid-level T33 into M55 with high percent choice, indicating reduced phonetic sensitivity in the categorisation process, but medium goodness ratings, indicating moderate residual sensitivity to phonetic deviations of M55 from T33 in the rating process. This finding is also in line with the claim from SLM (Flege, 1995) that a native equivalence classification to a native phoneme category can override detection of second language (L2) features (pitch height in this case) that are not contrastive in the native language, but that listeners can still retain sensitivity to phonetic differences in some tasks.

Additionally, the differing assimilation patterns for T45 between the Northern and Southern Vietnamese groups reflect native phonetic influences. PAM made different predictions for T45 assimilation by the Northern (Categorised as NV35) and Southern Vietnamese groups ($\text{UnCategorised}_{\text{clustered}}$ and split among SV35, SV214, SV21) as it considered the phonetic

differences between Northern Vietnamese and Southern Vietnamese rising/falling-rising tones. As can be seen in Figure 1, the phonetic trajectory and height of T45 is more similar to NV35 than to any other NV tones. On the other hand, T45 is roughly equivalently weak in similarity to the heights and trajectories of SV35, SV214, and SV21, the latter of which is realised with a slight final rise, unlike its Northern Vietnamese counterpart NV21. Native phonological system of Northern Vietnamese exhibited stronger phonological but weaker phonetic effects, resulting in Categorised assimilation, than that of Southern Vietnamese, resulting in $\text{UnCategorised}_{\text{clustered}}$ assimilation.

Also as predicted, we observed longer response times for $\text{UnCategorised}_{\text{clustered}}$ than for Categorised assimilations. We suggest that the Categorised cases indicate a strong, straightforward phonological influence from the native language tone system, where there is a single native category activated much more highly than other categories, with little to no phonological competition, resulting in fast responses. However, $\text{UnCategorised}_{\text{clustered}}$ assimilations reflect weaker native phonological influence and involve different processes from the Categorised ones. The longer response time in $\text{UnCategorised}_{\text{clustered}}$ assimilations likely reflects an extra processing cost and perceptual uncertainty, caused by competition among multiple weakly activated native categories and phonetic discrepancy from those native categories.

In summary, PAM accurately predicted assimilation types and/or percent choice and goodness rating ranges in the great majority of cases. We observed strong native phonological effects in Categorised assimilations, especially when percentage choice and goodness ratings were high. Strong phonetic effects reflected in residual sensitivity to phonetic variations within categories were indicated in low percent choice and/or goodness ratings; moderate phonetic effects were reflected in medium percent choice and/or ratings.

We acknowledge that PAM predictions typically focus on the relationships between assimilation patterns for non-native contrasts and discrimination levels for those contrasts. Discrimination was not addressed in the current study, which examined phonological and phonetic influences of different native tone systems on the listeners' assimilation patterns for unfamiliar Thai tones. We also wished to identify Thai tone contrasts for future research examining the PAM predictions for discrimination of tone contrasts based on the current tone categorisation findings. We examined the categorisation-discrimination relationship in a separate study (Chen, Best, Antoniou, et al., 2019) in which we found that Mandarin listeners had greater difficulties in discriminating T45 and T315, which were Categorised into the same native category (Single-Category assimilation), than in discriminating T21-T33, which were Categorised into different native categories (Two-Category assimilation). It would be important to look at discrimination of the T21-T33 contrast by Southern Vietnamese listeners, who instead Categorised it into the same native category. According to PAM principles, they should discriminate that contrast significantly more poorly than the Mandarin listeners did.

Findings of assimilation patterns not only carry theoretical implications for non-native perception research, but can also shed light on second language speech learning (Best, 2019). For example, according to PAM-L2 (Best & Tyler, 2007), when a non-native (L2) phone is Categorised as a good exemplar of a L1 phonological category, no further perceptual learning is likely to happen. This can be beneficial to L2 learning when the native category is phonetically similar to the non-native category. However, Categorised assimilation could be counterproductive in L2 tone production accuracy if the non-native category is phonetically different from the corresponding native category. The current study did not examine L2 tone production and its relationship to perceptual assimilation of tones, which also addresses the core principles of SLM (e.g., Flege,

1995). Future research could do so by investigating non-native tone imitation. For example, assimilating T33 into M55 could lead Mandarin learners of Thai to produce T33 with an inappropriately high F0. However, if Mandarin listeners perceive T33 as functionally equivalent to M55 at the phonological level but perceive a phonetic difference between L1 and L2 tones, they may learn to refine the phonetic details of their production for the L2 tone. In a preliminary imitation study using the same stimuli (Chen, Best, & Antoniou, 2019), we found that Mandarin participants' imitation of T33 was accurate in terms of F0 height, suggesting that while they perceptually Categorised T33 as M55, the phonetic differences between T33 and M55 were retained in memory and available for imitation and L2 speech learning.

For non-native tones that are not Categorised as any single L1 phonological category but are heard as being similar to several L1 categories), i.e., UnCategorised assimilations, PAM-L2 predicts that one or more new L2 phonological categories may be formed. Similarly, SLM claims that new categories could be formed for this type of L2 phone, and that if the new phonetic category matches that of native speakers of the L2, then the L2 sound will be produced accurately. However, PAM-L2 predictions differ from SLM in that PAM considers the comparative relationships within the interlanguage phonological system in addition to the similarity of a given L2 phone to the closest individual native phonetic category. If the UnCategorised L2 phones are assimilated into different sets of L1 phonemes with little overlap in the native categories chosen, PAM-L2 predicts that two or more new categories could be formed. But if the UnCategorised L2 phones are identified as similar to the same set of L1 phonemes, then only a single new category would be formed, and discrimination of the contrasting L2 phones may remain difficult (Best et al., 2019). Further research could examine the discrimination of non-native tone contrasts in which the two tones are assimilated to overlap or non-overlap set(s) of native tones by tone language listeners.

5.5 Conclusions

To conclude, our findings demonstrate that perceptual assimilation of non-native tones is affected by listeners' native phonological categories as well as the phonetic details of those categories. PAM predictions, which consider both phonological and phonetic factors in predicting non-native assimilation, were upheld. Naïve listeners Categorised many non-native tones into their native tone phonological categories while at the same time often retaining some degree of sensitivity to phonetic differences between the non-native and native tones. When there was only one corresponding native phonological category for a non-native tone, phonological constraints overrode phonetic differences in affecting percent choice in categorisation; however, listeners still retained residual phonetic sensitivity in goodness rating in some cases, indicated by low rating scores. Phonological and phonetic differences between two dialectal variants of Vietnamese affected assimilation types, reflecting native phonological effects, as well as percent choice and goodness-ratings, reflecting native phonetic effects. In addition, strong phonological influences facilitated Categorised assimilations, resulting in shorter categorisation response times than UnCategorised assimilations, which reflect weak native phonological influence and are more affected by non-native phonetic discrepancy and competition among alternative native categories.

The current findings have substantive implications for theories of tone perception as well as for second language lexical tone learning. Future research should consider the language-specific tone assimilation patterns found in the present study and compare discrimination performance among the same non-native tone contrasts that are assimilated differently by naïve listeners of different native language backgrounds. Furthermore, future research should also compare their imitations of non-native tones with their assimilations of those tones, to detect native language perceptual

influences on production, as hypothesised by SLM and anticipated by PAM assumptions about perception of articulatory information in speech.

Chapter 6. Cognitive factors in the perception of non-native lexical tones by tone language listeners

Juqiang Chen[a*], Catherine T. Best[a,b*], Mark Antoniou[a]

[a] Western Sydney University, The MARCS Institute for Brain Behaviour and Development, Penrith, NSW 2751, Australia

[b] Haskins Laboratories, New Haven CT, USA

6.1 Introduction

Perception of non-native phones is susceptible to native language influences at both phonological and phonetic levels. If a consonant or vowel contrast does not exist in the native language, listeners may have difficulties in identifying and discriminating it, e.g., native Japanese speakers have difficulties discriminating the English /r/-/l/ contrast which is absent in Japanese (MacKain et al., 1981; Miyawaki et al., 1975; Sheldon & Strange, 1982; Yamada & Tohkura, 1992). Even when a contrast in an unfamiliar language does exist in the listener's native system, however, differences in phonetic realisation of the native versus non-native phones can also affect perception (Best & Strange, 1992; Hallé, Best, & Levitt, 1999). For example, /w/-/r/ is phonologically contrastive in French but /r/ in French is a uvular approximant or trill and thus differs phonetically from English /r/. Consequently, when French listeners perceived an English /w/-/r/ continuum, they categorised /r/ less consistently than AE listeners. These effects, moreover, can be substantially modulated by memory load (Asano, 2017; Pisoni, 1973; Werker & Tees, 1984; Werker & Logan, 1985), talker variability (Magnuson & Nusbaum, 2007; Mullennix & Pisoni, 1990), and phonetic context variability (Shaw & Tyler, 2020; Zheng, 2014), which can shift listeners toward a more phonological mode of perception.

Most research has focused on perception of non-native consonants and vowels. Few studies have investigated how these factors affect non-native perception of lexical tones. One of those few found that both native phonological and phonetic effects on perceptual assimilation of Thai tones by Mandarin and Vietnamese listeners, but it did not examine discrimination (Chen et al., 2020). The present study made use of those assimilation results to predict how native listeners of Mandarin and Vietnamese would discriminate different types of Thai tone contrasts. In the next

section, speech perception theories will be reviewed regarding how they account for native language phonological and phonetic effects.

6.2 Native language phonological and phonetic effects on non-native perception

Several theoretical frameworks have been proposed to account for the influences of listeners' native language on non-native speech perception. As the focus of this paper is on the perception of non-native tone *contrasts*, the Speech Learning Model (SLM, Flege, 1995) and the Native Language Magnet model (NLM, Kuhl & Iverson, 1995), which focus more on individual phonetic categories rather than on phonological contrasts, will not be considered further. The Second Language Linguistic Perception model (L2LP: e.g., Escudero, 2005; Escudero & Vasiliev, 2011; Escudero & Williams, 2011) will also not be considered further here, as it has focused mainly on perceptual assimilation of vowels with predictions based on acoustic similarities, not considering phonological and phonetic overlap in detail. The Perceptual Assimilation Model (PAM, Best, 1995) was selected as the theoretical framework for the present study because it provides a coherent account of how native phonological and phonetic properties affect both perceptual assimilation of non-native phones into native categories, and how varying assimilation patterns influence discrimination of non-native contrasts.

PAM (Best, 1995) posits that naïve adult listeners perceptually assimilate non-native phones to native phonemes based on the perceived gestural (articulatory-phonetic) similarities between them. A non-native phone can be heard (1) as a good or poor exemplar of a native phoneme, i.e., Categorised, or (2) as less strongly like any single native phoneme but still falls within the native phonological space, i.e., UnCategorised or (3) as a non-speech sound, i.e., Non-Assimilated. When two phones of a non-native contrast are Categorised into two native categories, i.e., Two-Category assimilations, they are predicted to be better discriminated than when two non-native phones are

Categorised into the same native category but differ in their degrees of perceived discrepancy from the native "ideal", i.e., Category-Goodness assimilations. When two non-native phones are Categorised into the same native category as equally good to bad exemplars of the category, i.e., Single-Category assimilations, these contrasts are the most difficult to discriminate.

Perceived phonological overlap between contrasting non-native phones can decrease discrimination accuracy within Two-Category, UnCategorised-Categorised and UnCategorised-UnCategorised contrast assimilation types (Antoniou et al., 2013; Best et al., 2019; Faris et al., 2018; So & Best, 2014). *Non-overlap* contrast assimilations (e.g., Two-Category/Non-overlap, UnCategorised-Categorised/Non-overlap, UnCategorised-UnCategorised/Non-overlap) refer to the cases where non-native phones are each identified with completely different sets of native categories. When there are one or more shared above-chance categories for assimilation of the contrasting non-native phones, but not all choices are shared, the contrast assimilation is *partially overlapped* (e.g., Two-Category/Partial-overlap, UnCategorised-Categorised/Partial-overlap, UnCategorised-UnCategorised/Partial-overlap). Finally, when all the above-chance categories are the same for both non-native contrasts, the contrast is *completely overlapped* (e.g., Two-Category/Complete-overlap, UnCategorised-Categorised/Complete-overlap, UnCategorised-UnCategorised/Complete-overlap). *Non-overlapped* contrasts are facilitated by native phonological distinctions and should be better discriminated than *partially overlapped* contrasts which in turn should be better discriminated than *completely overlapped* contrasts, discrimination of which are interfered by native phonological similarity (Best et al., 2019).

Relatively few studies have investigated non-native tone perception and they often examined only categorisation (Chen et al., 2020; Wu et al., 2014) or only discrimination (Hao, 2017; Lee, Vakoch, & Wurm, 1996). Of the few that examined both categorisation and discrimination (Reid et al.,

2015; So & Best, 2014; Wu et al., 2015), some tested only non-tone language listeners (So & Best, 2014). For non-tone language speakers, only consonants and vowels, but not lexical tones, are segmental categories in their phonological systems (Best, 2019) and they may therefore assimilate non-native tones into their native intonation/prosodic categories rather than segmental categories. However, that type of assimilation is likely to differ qualitatively from assimilation to a corresponding native tone category by listeners of other tone languages, in which tones serve segmental-level functions (Duanmu, 1990, 1994; Lin, 1989). Thus, the relation between perceptual assimilation and discrimination of non-native tone contrasts could be different for listeners of other tone languages than for non-tone language listeners. Although some other studies have tested assimilation and discrimination of non-native tones by native tone language listeners, e.g., Mandarin listeners perceiving Thai tones (Reid et al., 2015) and Mandarin listeners perceiving Cantonese tones (Wu et al., 2015), they often only tested one native tone language group and thus were unable to verify their results across other tone languages. Thus, it remains unresolved how native phonological and phonetic factors contribute to non-native lexical tone perception.

6.3 Phonetic versus phonological mode of speech perception

Performance varies in online processing of continuous speech by adult naïve listeners and second language learners. The Automatic Selective Perception model (ASP, Strange, 2011) proposes that listeners can switch between a phonological and phonetic mode when perceiving native and non-native speech. When the attention focus of a task is on distinguishing phonetic differences that are essential to lexical distinctions, i.e., phonological distinctiveness (Best, 2015; Best et al., 2009), the phonological mode of speech perception is used. On the other hand, when the task requires listeners to attend to finer-grained phonetic details, the phonetic mode should be used to detect phonetic variations within native phonological categories.

In non-native perception, phonological selective perception routines that are attuned to the native phonological system are used automatically in the phonological mode, leading to the strongest native phonological influence. Consequently, phonetic differences within a native phonological category are less likely to be detected. However, in the phonetic mode, non-native phonetic differences that fall in a single native phonological category are more likely to be detected than when they are processed in a phonological mode.

Most studies that have examined different modes of perception in non-native speech perception so far have only used discrimination tasks but not categorisation tasks (Asano, 2017; Werker & Tees, 1984; Werker & Logan, 1985). Discrimination emphasises the detection of phonetic distinctions whereas categorisation involves accessing internalised phonological representations to identify the abstract category. Thus, categorisation task inherently requires more phonological level of processing whereas discrimination is more likely to rely on a more phonetic level of processing. It remains unresolved how listeners will alter between phonological and phonetic mode in discrimination and categorisation as a function of other factors that shift processing toward or away from concrete phonetic details. For example, the phonetic mode of perception depends more on the availability of phonetic details in short-term memory than the phonological mode of perception and thus is susceptible to memory load effects as well as to phonetic variability in speech signal.

6.3.1 Memory load

The availability of phonetic details in short-term memory determines listeners' ability to use a phonetic mode of perception. Listeners can only retain the rich array of fine-grained phonetic details in short-term memory for a limited time before they rapidly decay (Baddeley, 2010; Baddeley & Hitch, 1974). We refer to the amount of time that listeners must wait before making

final decisions as memory load. The longer the interval between two stimuli or between stimulus and response, i.e., high memory load, the more likely memory of the full range of phonetic details will have faded by the time the perceptual decision is made. As a result, listeners will use a phonological mode of perception.

Consistent with that analysis, several studies have found that longer intervals between stimuli (Interstimulus Interval, ISI) lead to poorer discrimination of "difficult" non-native consonant contrasts (Asano, 2017; Werker & Tees, 1984; Werker & Logan, 1985) that appear to fit the PAM's definition of Single-Category assimilations, i.e., perceived as equivalent phonetic variants within a single native category. Studies on discrimination of non-native consonants have reported a switch from phonetic to phonological mode as a function of memory load. In one pioneering study on this issue, English participants discriminated the "difficult" non-native Hindi voiceless unaspirated retroflex versus dental stop consonant contrast under two memory loads. They discriminated the contrast under low memory load (ISI = 500 ms), but not under high memory load (ISI = 1500 ms) (Werker & Logan, 1985). Similarly, the performance of German native listeners decreased significantly when discriminating non-native Japanese consonant length contrasts under high memory load, relative to low memory load (Asano, 2017).

Unlike perception of non-native consonants, the memory load effect appears to be different in tone perception. For example, when Mandarin listeners were asked to discriminate Cantonese tones or Cantonese listeners discriminated Mandarin tones, their very good performance was unchanged even under greatly lengthened ISI conditions (5 s) (Lee et al., 1996). Similarly, Yu and colleagues (2017) used a passive oddball paradigm with short vs. long ISIs (600 vs. 2600 ms) to test discrimination of Mandarin tones by native listeners and naïve English listeners who had no exposure to tone languages. Both behavioural and event-related brain potential data indicated no

main effect of ISI for either the native Mandarin or the naïve English listeners, although the English group showed significantly lower accuracy than the Mandarin group in the long ISI condition. According to the cue-duration hypothesis (Fujisaki & Kawashima, 1970), the acoustic cues responsible for distinguishing consonants, e.g., formant transitions or voice onset time, are generally short in duration and transient, and cannot be retained in detailed form in memory, whereas the phonetic details of longer-duration speech segments such as vowels can be better retained in memory. Thus, consonants are more likely to suffer from rapid memory decay than vowels. As the acoustic cues of lexical tones extend over the whole sonorant portion of a syllable, i.e., the syllable's fundamental frequency contour, we argue that tones, like vowels, will also be retained in memory fairly well over time. Nonetheless, it remains unknown whether the pattern observed by Lee and colleagues' (1996) with Cantonese and Mandarin tones, which are both dialects of Chinese, will be seen when the target and listener languages are unrelated non-Chinese tone languages with different tone inventories, such as Thai as perceived by Mandarin and Vietnamese listeners.

Extending the idea of fading phonetic details in short-term memory to categorisation, longer delay in responding to a target stimulus in categorisation tasks should lead to decay of phonetic details in short-term memory, shifting listeners to a phonological mode of perception. However, this has not been systematically examined in prior studies of non-native tone categorisation by tone-language listeners, nor indeed in studies of non-native consonant or vowel categorisation. We hypothesise that in categorisation, the phonetic details of a non-native tone are initially stored in working memory. During this time, the phonological categories of the listeners' native tones that are most phonetically similar to the non-native tone are activated and retrieved from long-term memory to be compared with the incoming non-native tone. After having heard the target tone,

the longer listeners must hold its details in working memory, the more its phonetic trace fades away, increasing the need for participants to rely on their native phonological categories rather than on the faded phonetic details of the target stimulus. Thus, longer response intervals in categorisation will increase memory load and induce stronger native phonological influences. We propose a gradient that indicates the strength of native phonological influence in perceptual assimilation: Categorised > $\text{UnCategorised}_{\text{focalised}}$ > $\text{UnCategorised}_{\text{clustered}}$ > $\text{UnCategorised}_{\text{dispersed}}$ assimilations. That is, Categorised assimilation indicates stronger native phonological influence than UnCategorised assimilation. A Categorised non-native phone under high memory load, indicating strong native phonological influence, could change to be UnCategorised under low memory load, which indicates weaker native phonological influence.

Phonetically, listeners retain some residual sensitivity to within-category phonetic variations from their native categories, which is reflected in the percent choice and goodness ratings in categorisation tasks (Chen et al., 2020). High residual sensitivity is reflected in low percent choice and category-goodness ratings whereas low residual sensitivity is reflected in high percent choice and ratings. When memory load is low and phonetic details are available in short-term memory, listeners will show high residual sensitivity to within category phonetic variations. When memory load is high and phonetic details have faded, listeners will show low residual sensitivity instead.

6.3.2 Stimulus variability: Talker and phonetic context differences

Extending principles from ASP, when dealing with substantial linguistically irrelevant variability in speech, listeners switch to a phonological mode of perception where they perceive abstract phonological information from the speech because phonetic information is variable. On the other hand, listeners shift to a phonetic mode of perception when variability of speech is low because

phonetic information is constant and reliable in this case. Two most common types of variability are talker and phonetic context variabilities. Two talkers produce the same segment with different acoustic-phonetic details due to differences in the shape and length of their vocal tracts and other indexical characteristics that affect speech production (Nusbaum & Morin, 1992). Listeners need to somehow discount such linguistically irrelevant talker-specific information to perceive constant phonological categories across different talkers. Processing talker variability affects speech perception: poorer performance in consonant and vowel recognition (Nusbaum & Morin, 1992), identification (Mullennix et al., 1989), tone identification (Wiener & Lee, 2020; Wong & Diehl, 2003).

Similarly, varying phonetic contexts of a target phoneme also lead to phonetic variations in that phoneme and consequently affect its perception. Lexical tones are affected by vowel contexts. Other things being equal, high vowels, such as [i], have a higher intrinsic F0 than low vowels, such as [a], in a number of languages (Ewan, 1975; Whalen & Levitt, 1995). Consistent with these intrinsic F0 effects, native tone language speakers' perception of tones is affected by the vowel context (Shaw & Tyler, 2020; Zheng, 2014). However, it remains unknown whether varying vowel variability will affect discrimination and categorisation of non-native tones relative to a constant phonetic context.

It is hypothesized that high stimulus variability should lead listeners to use a phonological mode of perception and rely more on abstract phonological information rather than on specific phonetic details in order to accomplish perception tasks. In perceptual assimilation, their percent choice and goodness ratings should be higher, reflecting low residual phonetic sensitivity to within category phonetic variations in high variability conditions. On the other hand, in discrimination, accuracy should be reduced especially for non-native contrasts within the same native phonological category

in high variability conditions, such as those that are assimilated as Category-goodness or Single-category contrasts.

6.4 Lexical tones in Thai, Mandarin and Vietnamese

We relied on existing Chao transcriptions to provide an initial approximation of the phonetic characteristics of the tones in our selected languages. We note, however, that Chao transcriptions are nonetheless based on perceived pitch within speakers' vocal range, in which 5 represents the highest and 1 the lowest pitch in that range (Chao, 1930). Here we distinguished between phonological features, which include perceived abstract pitch contours, i.e., level, rising, falling, rising-falling, falling-rising and heights, i.e., high, mid, low, versus specific, concrete F0 properties as show in 6.1. Thai (T) contrasts five lexical tones as shown in the left panel of Figure 6.1: three phonologically level tones, high-level, mid-level and low-level, and two contour tones, rising and falling (Gandour, 1978). Two of the three phonologically level tones in Thai are not *phonetically* flat, as can be seen in Figure 1: high-level is phonetically characterised as T45 in Chao transcription, and low-level as T21, while only mid-level T33 is phonetically flat (Reid et al., 2015). When produced in isolation, the phonologically rising tone is phonetically transcribed as T315 with a falling portion at its onset and the phonologically falling tone as T241 with a short rise at its onset.

Mandarin (M) has four tones (see Figure 6.1, middle panel), of which one is phonologically level and three are phonologically contoured in their citation forms: a high-level tone M55 (called Tone 1 in Mandarin); a rising tone M35 (Tone 2); a falling-rising tone M214 (Tone 3); and a falling tone M51 (Tone 4) (Chao, 1968). Vietnamese[18] has five tones in its phonological system (Figure 6.1,

[18] our Vietnamese listeners all spoke the Southern Vietnamese dialect.

right panel): two phonologically level tones, high-level V44, also called *ngang* in Vietnamese, and low-level V22 (*huyền*); and three phonologically contour tones: rising V35 (*sắc*), falling V21 (*nặng)*, which appears to have a modest final rise in Figure 6.1, and falling-rising V214 (Nhàn, 1984). V214 results from a merger of the two falling-rising tones in the standard/Northern dialect, NV214 (*hỏi*), and NV415 (*ngã*) (Brunelle, 2009).

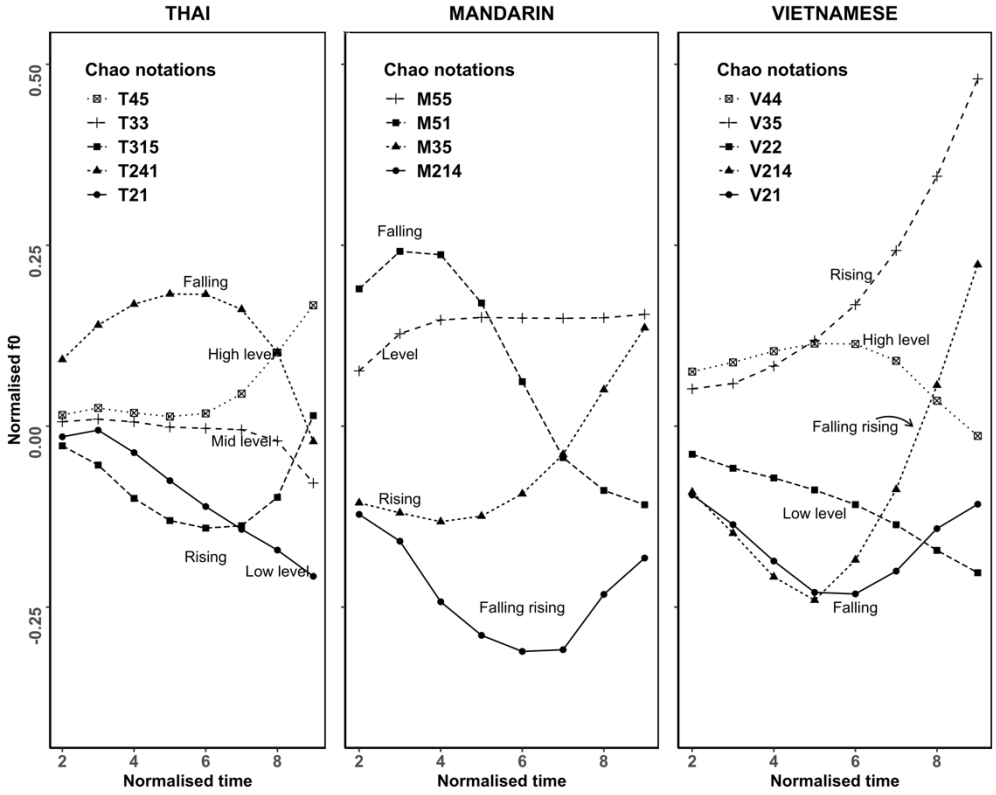


Figure 6.1 Time- and Lobanov-normalised (Lobanov, 1971) F0 contours of Thai, Mandarin and Southern Vietnamese tones[19]. The legends in each panel show the Chao notations for the tones of the respective languages.

[19] in Chen et al (2020), Southern Vietnamese speakers produced both V214 (hỏi) and V415 (ngã) with no significant acoustic differences, consistent with the reports of merger, so here they were averaged and labelled as a single phonologically falling rising tone SV214.

In this paper, we report two experiments that examined phonological and phonetic influence of Mandarin and Vietnamese listeners' native tone systems on their perception of Thai tones and investigated how memory load and stimulus variability bias listeners to a phonological versus phonetic model of perception in discrimination and categorisation tasks.

6.5 Experiment 1: Mandarin listeners' perception of Thai tones

Experiment 1 was designed to examine perceptual assimilation and discrimination of Thai tones by Mandarin listeners. 500 ms was selected to be the interval under low memory load to ensure the availability of phonetic details in working memory (Asano, 2017) whereas 2000 ms was selected to be the interval under high memory load to maximise the decay of phonetic details (c.f., 1500 ms in Werker & Tees, 1984; Werker & Logan, 1985) and at the same time keep the duration of the experiment reasonable. Although a within-subjects design with blocked memory load conditions would grant greater statistical power in data analysis, participants are unlikely to switch between two modes of perception within an experiment, and if required to do so, for the second part of the study, they rely on strategies developed in the prior test blocks (Werker & Logan, 1985). For this reason, a between-subjects design was employed for the memory load conditions.

Five Thai contrasts were used in the discrimination task, selected based on findings from a previous study with these stimulus and listener languages (Chen et al., 2020) to present a rich range of assimilation patterns and predicted differences between the two listener languages. In that study, Mandarin listeners Categorised T45 to their native M35, T33 to M55, T21 to M214 and T315 to M35. T241 was an UnCategorised assimilation split between M55 and M51. For the present study, we selected five Thai contrasts T241-T21, T33-T21, T315-T45, T33-T241 and T33-T45. In the prior study T241-T21 had formed an UnCategorised-Categorised assimilation for Mandarin listeners; T33-T21 a Two-Category assimilation; T315-T45 an Single-Category assimilation; T33-T241 an UnCategorised-Categorised assimilation; and T33-T45 a Two-Category assimilation. However, in that study the categorisation response intervals were not manipulated, and participants were instructed simply to respond as quickly as possible. As we manipulated the response intervals

in the categorisation task in the present study, the assimilation results could differ somewhat from the untimed responses of the prior study. We therefore based our discrimination predictions for the current study on the assimilation results we obtained here with manipulations of the response interval.

6.5.1 Method

6.5.1.1 Participants

32 native speakers of Mandarin participated and were divided equally into two groups for each memory load condition (see *Procedure*) (Low: M_{age} = 26.6 years, Age range: 18-48, 10 females, 6 males; High: M_{age} = 26.0 years, Age range: 18-39, 10 females, 6 males). The Mandarin-speaking participants were all born and raised in various regions in China (i.e., Henan, Hunan, Jilin, Jiangsu, Jiangxi)[20] but were educated in Mandarin from early childhood through high school, and they used Mandarin on a daily basis. None of them spoke Cantonese. None had more than two years of formal musical training, which is known to influence tone perception (Gottfried et al., 2004). All reported normal hearing, and none had experience with Thai. The experiments were approved by the Western Sydney University Human Research Ethics Committee (approval H12560) and all participants gave informed consent form prior to testing.

6.5.1.2 Stimulus materials

Two syllables (/ma/, /mi/) were chosen for target stimuli because they are legal and form

[20] In future research, it would be desirable to form more homogeneous groups of participants and have a better control of their dialects, e.g., recruiting only Beijing Mandarin speakers. For practical reasons, strict control of dialect background of Mandarin participants in Sydney is difficult if not impossible. With this said, the possible effect of dialect background differences adds potential variation to the assimilation patterns. In experiments in Chapters 6 and 7, we used participants' assimilation patterns to directly predict and account for variations in discrimination and imitation accuracy, rather than using phonetic descriptions of any Mandarin dialect variations. In this way, dialect effects should be consistent across assimilation, discrimination, and imitation tasks.

meaningful morphemes for each native tone in Thai and Mandarin. Thus, naïve listeners were able to categorise the Thai tone stimuli into their native phonological systems as morphemes. The Thai syllables were recorded for a separate study (Burnham et al., 2009), and were used here with permission from the authors. The target Thai syllables were each read several times by two female native Thai speakers who had no experience with other tone languages. Two tokens of each target item per speaker were selected and were judged to be correct and most natural sounding to a third native Thai speaker.

In all the following experiments, we systematically manipulated two sources of stimulus variability: number of talkers (one vs two) and/or vowel contexts (one versus two). Tone realisation is affected both by F0/contour differences between talkers and by intrinsic F0 differences between vowels, which is higher for the high vowel [i] in *mi* than in the low vowel [a] in *ma* (see Figure 6.2).

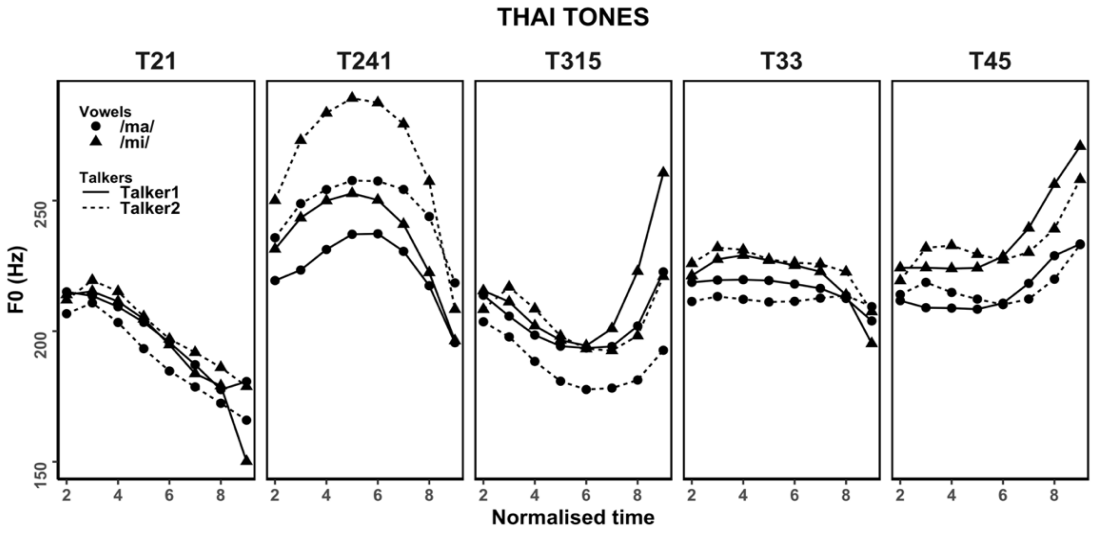


Figure 6.2 Time-normalised F0 contours of the five Thai tones produced by two female informants with /ma/ and /mi/ (each with five tokens, including the two tokens per target per speaker that were used in the perceptual experiment).

6.5.1.3 Procedure

Participants were tested individually in a quiet testing space. Stimuli were presented on a Dell

Latitude 7280 laptop running E-Prime Professional 2, via Sennheiser HD 280 Pro headphones at 72 dB SPL. In all experiments, discrimination tests were run before categorisation tests to minimise the influence of categorisation on discrimination decisions. Two groups of Mandarin participants took part in Experiment 1 with level of memory load held constant within each group across the two tasks. The low memory load condition used 500 ms as the interstimulus interval (ISI) in the AX discrimination task and as the Response Interval (RI) in the categorisation task; the high memory load condition used 2000 ms ISIs and RIs, respectively.

Discrimination

An AX discrimination task ("same-different") was used because it allows for a single ISI interval between the stimuli to be judged, thus offering a clearer interpretation of memory load effects than tasks such as AXB or 4AFC which respectively require two or three ISI intervals on each trial. The AX trial format has been used in many previous discrimination studies involving ISI manipulations (Asano, 2017; Werker & Tees, 1984; Werker & Logan, 1985). On each trial, participants heard a stimulus A followed by a silence of either 500 or 2000 ms. After the ISI, the second stimulus X was presented and "Same or different?" appeared on the screen to signal to the participant that they should respond. They were told to only consider the tone. Half of the trials were "same" trials and the other half were "different" trials; these were presented in random order. Participants were given 3 s to answer before timeout on each trial, and the inter-trial interval following their response (or timeout) was 1 s. Trials with time-outs were missing data, which accounted for 0.5% of all the data across all Mandarin participants in this task. In each memory load condition, both talker and vowel variability conditions, i.e., Constant vs Variable: one versus two talker/vowel contexts, respectively, were manipulated across eight randomised blocks (Table 6.1). Each of the resulting stimulus pairings was repeated two times in four AX trial types, i.e., for

stimulus A and B: AA, AB, BA, BB, randomly within each testing block. Before the discrimination test, participants completed 16 practice trials with same syllables produced by a third Thai speaker which were not used in the main experiment.

Table 6.1 Stimulus details for the AX discrimination task in talker and vowel variability conditions. (40 trials per block)

Conditions	Details
Constant Talker + Vowel	Same Talker, same Vowel in each of 4 blocks: Talker 1 /mi/; Talker 1 /ma/; Talker 2 /mi/, Talker 2 /ma/
Constant Talker/ Variable Vowels	Same Talker, both Vowels in each of 4 blocks: Talker 1 /ma/→/mi/; Talker 1 /mi/→/ma/; Talker 2 /ma/→/mi/; Talker 1 /mi/→/ma/
Variable Talkers/ Constant Vowel	Both Talkers, same Vowel in each of 4 blocks: Talker 1→ Talker 2 /ma/; Talker 2→ Talker 1 /ma/; Talker 1→ Talker 2 /mi/; Talker 2→ Talker 1 /mi/
Variable Talkers + Vowels	Both Talkers, both Vowels in each of 4 blocks: Talker 1 /ma/ → Talker 2 /mi/; Talker 1 /mi/ → Talker 2 /ma/; Talker 2 /ma/ → Talker 1 /mi/; Talker 2 /mi/ → Talker 1 /ma/

Categorisation

On each trial, a stimulus token was presented, and listeners made a forced-choice categorisation judgment to their native tones in four Pinyin options via a keypress within a 3 s time limit. The stickers were placed on the keys in the same line on the keyboard, i.e., “f”, “g”, “h”, “j” for Mandarin participants, “f”, “g”, “h”, “j”, “k” for Southern Vietnamese participants. None of the

participants reported any problems using these keys which were handy when participants were familiar with their positions. In addition, these keys are quite close to each other, and only key responses but not reaction times were analysed and thus the positions of response keys would have had little impact on the results. They were instructed to press the key only after they saw "which tone?" onscreen. The interval between stimulus presentation and the signal to answer ("which tone?"), i.e., response interval or RI, was 500 ms under low memory load and 2000 ms under high memory load in line with our manipulation of ISI in the discrimination task. We note that the signal to respond, "Which tone?", in the present experiment did not actively prevent listeners from deciding tone response categories in their mind before the signal, but it did prevent them from *responding* until they received the signal to go. Immediately after their response, they heard the tone again and rated goodness of fit into their chosen native category on a 7-point scale via a keypress: 1 = poor, 7 = perfect. Talker and vowel variability were manipulated across nine blocks (see Table 6.2). Each tone was tested under each stimulus variability condition with two tokens from each speaker except for the variable-talker-variable-vowel block, in which ten /ma/ trials (5 tones × 2 talker) and ten /mi/ trials (5 tones × 2 talker) from one token were included, to avoid participant fatigue and make the experiment feasible in terms of time. The differences between two tokens of the same talker and vowel should be relatively small compared to talker and vowel variations. However, this condition had a smaller number of data points (n = 20) relative to other conditions (n = 40), which reduced some of its statistical power but for the assimilation task only. In addition, this should not impact the memory load effect on assimilation, as participants in both groups did the same task. Before the categorisation test, participants completed 10 practice trials with same syllables produced by a third Thai talker which were also used for the AX practice trials but not used in the main experiment.

Table 6.2 Stimulus details for the categorisation task in talker and vowel variability conditions. T = Talker(s); V = Vowel(s)

Conditions	Details
Constant Talker + Vowel	Same Talker, same Vowel in each of 4 blocks: Talker 1 /mi/ randomised (10 trials) + Talker 1 /ma/ randomised (10 trials); Talker 2 /mi/ randomised (10 trials) + Talker 2 /ma/ randomised (10 trials)
Constant Talker / Variable Vowels	Same Talker, both Vowels in each of 2 blocks: Talker 1 with /ma/ + /mi/ randomised (20 trials); Talker 2 with /ma/ + /mi/ randomised (20 trials)
Variable Talkers / Constant Vowel	Both Talkers, same Vowel in each of 2 blocks: /ma/ with Talker 1 + Talker 2 randomised (20 trials); /mi/ with Talker 1 + Talker 2 randomised (20 trials)
Variable Talkers + Vowels	All talker + vowel combinations presented randomly within a block (20 trials)

6.5.2 Results

The categorisation results are presented first because they are used to determine the assimilation types for each of the five Thai tone contrasts and form the basis for PAM predictions about discrimination.

6.5.2.1 Categorisation

Table 6.3 shows the mean percent choice of categorisations to Mandarin tones and goodness of fit ratings for each Thai tone. Missing data, i.e., responses before the signal “Which tone?” or responses after timeout, accounted for 0.5% of all the data and were deleted before analysis. Although in some studies on consonant and vowel perception, Categorised assimilation has been defined as the selection of one native category (Tyler et al., 2014) for a given non-native phone

above a fixed threshold (e.g., 70%), we followed two statistically-based criteria used in So and Best (2014), which are more sensitive to variations among different native tone systems than a fixed threshold would be, as it considers the number of tone categories in the listener's language. First, a given native tone must be selected significantly above chance level: 25% for Mandarin listeners with four native tone choices. Second, that same native tone category must be chosen significantly more often than any other response categories.

Table 6.3 Assimilation of Thai tones into Mandarin tone categories under low versus high memory loads. Categories in bold are choices that were significantly above chance: 25% for Mandarin; "*" = Categorised tone. Assimilations: C = Categorised, U = UnCategorised. Rating: 1 = poor, 7 = perfect; mean ratings are displayed. "-" = no response.

Thai stimulus		T45		T33		T21		T315		T241	
	Response	%	rating	%	rating	%	rating	%	rating	%	rating
Low Memory Load	M55	0.4	4	**77.3***	5.3	19.9	3.3	-	-	21.6	4.2
	M35	**88.8***	5.8	2.8	2.8	1.6	2.1	**48.6**	5.5	2.9	3.7
	M214	10.3	4.3	0.7	3.3	25.8	3.7	**51.2**	5.4	0.4	3.5
	M51	0.4	5.5	19.2	4.9	**52.7***	4.6	0.2	7	**75.1***	5.6
Assimilation		C		C		C		$U_{clustered}$		C	
High Memory Load	M55	-	-	**84.7***	5.2	26.4	3.2	0.2	7	**28.1**	5
	M35	**85***	5.3	0.2	5	1.1	3.3	**44.2**	5.1	0.2	2
	M214	14.7	4.8	1.7	5.9	6.3	3.8	**55.6**	5.5	0.5	6
	M51	0.2	2	13.4	4.4	**66.2***	4	-	-	**71.2***	5.1
Assimilation		C		C		C		$U_{clustered}$		C	

We used *t*-tests to determine whether the response categories for each Thai tone exceeded chance level (> 25%). All above chance level choices were statistically significant, except for choices of M214 for T21 under low memory load and choices of M55 for T21 under high memory load (see

Appendix B, Table B.1 for the full set of statistical details).

In order to determine whether the above-chance-level response categories were chosen significantly more often than other response categories, we fitted the data using a Linear Mixed Effect Regression (LMER) model with percent choice as the dependent variable, native categories as a fixed factor, and subject as the random intercept. To calculate the p-values for the fixed effects, we used the Kenward-Roger approximation to the degrees of freedom, as recommend by Halekoh and Hojsgaard (2014), and the *Anova* function from the *car* package (Fox & Weisberg, 2019) in *R* (R Core Team, 2018), with test specified as "F". We then ran multiple comparisons between native categories with the R-package *lsmeans* (Lenth, 2016). Following the above procedure, four Thai tones were deemed to have been Categorised to Mandarin tone categories by our Mandarin listeners: T21 was Categorised as M51; T33 as M55; T45 as M35; and T241 as M51 (for statistical results see Appendix B, Table B.2 and Table B.3). T315 assimilation was UnCategorised and split between M35 and M214.

In order to test the effects of cognitive factors, we fitted the data with the *multinom* function[21] in the R package *nnet* (Venables & Ripley, 2002). A full multinomial regression model of the categorisation data was conducted with Mandarin tone choices as a dependent measure, and memory load (Low and High), talker and vowel variability (Constant vs Variable) and Thai tones (T21, T33, T45, T241, T315) as fixed factors. Four additional models were built each with one

[21] This function does not cover mixed effects for subjects as a random factor. To incorporate subjects as a random factor we also ran a Bayesian Markov Chain Monte Carlo (MCMC) analysis using the *MCMCglmm* package (Hadfield, 2010). The p values for significant effects are pMCMC values. That model showed similar results as the multinomial model, i.e., a significant effect for ISI and tone types but not for talker and vowel variability. However, the result was not stable even with a large number of iterations due to the random sampling aspect of the method. To be consistent with later models, we report the more stable multinomial model results here.

fixed factor subtracted and these models were compared to the full model to determine the effects of cognitive load. Both memory load, $\chi^2(60) = 146.14$, $p < .001$, and Thai tone, $\chi^2(96) = 6342.04$, $p < .001$, showed significant effects in Likelihood ratio tests, suggesting that participants responded differently for different Thai tones and in different memory load conditions. However, the talker and vowel variability effects were non-significant.

6.5.2.2 Predictions for discrimination

The original PAM principles (Best, 1995), without consideration of overlap in assimilated categories, stated that discrimination of a Two-Category contrast should be excellent and that of a UnCategorised-Categorised contrast should be very good because native phonological distinctions, i.e., either a native phonological contrast or a distinction between a native category and not-that-category, assist with the discrimination. On the other hand, discrimination of a Category-Goodness contrast rests solely on within-category phonetic sensitivity without assistance by a native phonological difference and should fall anywhere between moderate to good depending on difference in perceived goodness of fit of each non-native phone to the native category they were assimilated to. More recently, it has been recognised that the degree of overlap in native category assimilations between contrasting non-native phones must also be considered (Best et al., 2019; Faris et al 2016, 2018; Tyler et al, 2014). Overlap of native categories selected above chance level reflects native phonological contributions whereas overlap in choices of native categories below chance level reflects native phonetic contributions. When two non-native phones are perceived as phonologically similar to the same set of native categories selected above chance level, i.e., complete overlap, discrimination of this contrast should be strongly interfered by native phonological similarity and be poorer than when a non-native contrast is assimilated to a similar but not completely same set of native categories, i.e., partial overlap. The discrimination of a partial overlap contrast should still be poorer than when the members of a non-native contrast are

assimilated to completely different sets of native categories, i.e., non-overlap, in which native phonological distinctions facilitate discrimination. Therefore, discrimination of an UnCategorised-Categorised/Partial-overlap contrast should be comparable to Two-Category/Partial-overlap but poorer than that of Two-Category/Non-overlap or Category-Goodness/Non-overlap.

First, we identified phonological assimilation types and overlap types for the Mandarin group based on perceptual assimilation results. The manipulation of memory load in the categorisation task affected assimilation and overlap types. T33-T45 and T33-T21 were both Two-Category contrast assimilations with no overlap (Two-Category/Non-overlap) under both memory loads, so it was predicted that discrimination of these contrasts would be equally excellent. T33-T241 was a Two-Category assimilation but with partial overlap (Two-Category/Partial-overlap) under high memory load. Under low memory load, T33-241 remained a Two-Category/Non-overlap assimilation. Thus, T33-T45 and T33-T21 should show the highest discrimination in both memory load conditions, while under low memory load only, these two contrasts should be more accurately discriminated than Two-Category/Partial-overlap contrast T33-T241.

Given that both T21 and T241 were assimilated to M51, we ran *t*-tests for each memory load condition to determine whether the contrast assimilation was a Single-Category or Category-Goodness type. The rating differences were significant, i.e., low memory load: $t(25) = -2.47$, $p = .02$; high memory load, $t(28) = -2.67$, $p = .01$, indicating Category-Goodness assimilation in both memory load. Thus, we did not find any Single-Category contrast assimilations in this experiment. T315-T45 was an UnCategorised-Categorised/Partial-overlap assimilation with one overlap in Mandarin categories chosen in both memory load conditions.

To indicate the differences between two memory load conditions, we provide two sets of PAM assimilation type/overlap type predictions accordingly, one per memory load condition. Under low

memory load, discrimination accuracy predictions are, from good to poor: {Two-Category/Non-overlap: T33-T45 = T33-T21 = T33-T241} > {Category-Goodness/Non-overlap: T241-T21} > {UnCategorised-Categorised/Partial-overlap: T315-T45}. On the other hand, under high memory load, T33-T241 became Two-Category/Partial-overlap. Consequently, discrimination accuracy from good to poor was: {Two-Category/Non-overlap: T33-T45 = T33-T21} > {Category-Goodness/Non-overlap: T241-T21} > {Two-Category/Partial-overlap: T33-T241} = {UnCategorised-Categorised/Partial-overlap: T315-T45}.

In addition, for non-native contrasts of the same assimilation and phonological overlap type, listeners' phonetic sensitivity to contrasts can vary as a function of perceived phonetic similarities of the contrasting tones, as indicated by percent choice and goodness ratings. For example, the Two-Category/Non-overlap contrasts, such as T33-T45 and T33-T21, can differ in their perceived phonetic overlap, although their phonological overlap types, calculated on only above-chance native category choices, were the same. One approach for considering this phonetic effect is by computing the *overlap scores* (as in Flege & MacKay, 2004; Levy, 2009), calculated as the summed percentages of all overlapping native category choices, including below chance choices, for a non-native contrast. Goodness ratings were not considered in those previously-developed overlap scores, however. Given that both goodness ratings and percent choice indicate residual phonetic sensitivity to variations of the non-native stimulus from the chosen native category (Chen et al., 2020), we also calculated a *fit-index* (Wu et al., 2014) by multiplying percent choice by goodness ratings (see Appendix B, Table B.7). A high overlap score between contrasting non-native phones indicates *low* listener sensitivity to phonetic differences between them. Conversely, a high fit index difference between contrasting non-native stimuli, or *the fit index difference score*, represents high residual phonetic sensitivity to differences between them.

In the present study, we used the confidence intervals around the overlap scores and fit-index difference scores of contrasts with the same assimilation and/or overlap type to indicate their phonetic overlap degrees and predict differences in their discrimination. For example, T33-T45 and T33-T21 were both assimilated to Mandarin tones as Two-Category/Non-overlap. The overlap scores were lower, and the fit indices were higher for T33-T45 ($M_{\text{Overlap\_score}}$ = .039/.009; $M_{\text{Fit\_index}}$ = 1.75/1.76) than for T33-T21 ($M_{\text{Overlap\_score}}$ = .36/.41; $M_{\text{Fit\_index}}$ = 1.07/.99) under both memory loads. By the reasoning we provided above, discrimination should thus be better for T33-T45 than T33-T21. This results in a change to just one prediction based on assimilation and overlap type, which applies under both high and low memory load. Rather than discrimination performance being equivalent for the two contrasts, T33-T45 = T33-T21, as predicted earlier, differences in residual phonetic sensitivity for these two non-native tone contrasts should result in differential discrimination: T33-T45 > T33-T21. All other predictions listed earlier remain unchanged.

6.5.2.3 Discrimination

In order to minimise response bias effects on the AX data, we calculated *d′* (Macmillan & Creelman, 2005) for discrimination of each tone pair in each cognitive condition, with adjustments made for probabilities of 0 and 1 to avoid infinite values. The strategy is to add 0.5 to all raw data cells regardless of whether zeroes are present (Hautus, 1995; Miller, 1996). Hit is defined as the number of correct "different" responses on AB or BA trials. False positive is defined as the number of incorrect "different" responses on AA or BB trials. The *d′* scores were calculated using a differencing rule for AX tasks (as in Macmillan & Creelman, 2005) for each participant and separately for each tone pair and for each block. We fitted the data using an LMER model with *d′* as the dependent variable, and memory load (Low and High), talker and vowel variability (Constant vs Variable) and tone contrasts (T241-T21, T33-T21, T315-T45, T33-T241, T33-T45) as fixed factors, and participant as a random factor including random slopes for within-subject

fixed factors, i.e., talker and vowel variability conditions (as suggested by Barr et al., 2013). To calculate the *p*-values for the fixed effects, we again used the Kenward-Roger degrees of freedom approximation (Halekoh and Hojsgaard, 2014), and the *Anova* function from the *car* package in R, with test specified as "F" (see Table 6.4, for statistical estimates see Appendix B, Table B.9).

Table 6.4 Main effects and interactions for Thai tone contrast discrimination by Mandarin listeners. "*" indicates significant.

Cognitive factors	F	*df*		*p*
Memory load	0.08	1	30	0.78
Talker variability	41.08	1	30	< .01*
Vowel variability	57.40	1	30	< .01*
Tone contrasts	127.23	4	510	< .01*
Memory load × Talker variability	0.04	1	30	0.84
Memory load × Vowel variability	0.02	1	30	0.88
Talker variability × Vowel variability	20.06	1	510	< .01*
Memory load × Tone contrasts	0.31	4	510	0.87
Talker variability × Tone contrasts	1.55	4	510	0.19
Vowel variability × Tone contrasts	1.03	4	510	0.39
Memory load × Talker variability × Vowel variability	0.41	1	510	0.52
Memory load × Talker variability × Tone contrasts	0.05	4	510	1.00
Memory load × Vowel variability × Tone contrasts	0.91	4	510	0.46
Talker variability × Vowel variability × Tone contrasts	0.33	4	510	0.86
Memory load × Talker variability × Vowel variability × Tone contrasts	0.12	4	510	0.98

There were significant main effects of talker variability, vowel variability, and Thai tone contrasts. However, the main effect of memory load and all interactions involving memory load were non-significant (see Table 6.4 for *p*-values of fixed effects and Table B.9 for estimates of the model). The scores for the constant-talker blocks (M = 3.26, 95% CIs [3.11, 3.41]) were significantly

higher, suggesting better discrimination, than those of the variable-talker blocks (M = 2.79, 95% CIs [2.64, 2.94]). Similarly, the scores for constant-vowel blocks (M = 3.11, 95% CIs [3.16, 3.46]) were significantly higher than those of variable-vowel blocks (M = 2.74, 95% CIs [2.59, 2.89]).

To further examine the Thai tone contrast main effect and test PAM predictions, we ran pairwise multiple comparisons with Tukey adjustment for each tone pair contrast (see appendix B Table B.10 for full statistical details). All contrast comparisons were significant, except for the T315-T45 versus T33-T241 comparison. These results reflect the following pattern: Two-Category/Non-overlap (T33-T45, M= 4.33, 95% CIs [4.15, 4.52]) > Two-Category/Non-overlap (T33-T21, M= 3.42, 95% CIs [3.21, 3.63]) > Category-Goodness/Non-overlap (T241-T21, M= 3.05, 95% CIs [2.85, 3.26]) > Two-Category/Non-overlap vs. partial-overlap (T33-241, M= 2.06, 95% CIs [1.87, 2.25]) = UnCategorised-Categorised/Partial-overlap (T315-T45, M= 2.26, 95% CIs [2.06, 2.45]).

In addition to main effects, multiple comparisons were conducted with Tukey adjustments to break down the talker × vowel interaction (see Figure 6.3). In the constant-talker condition, the constant-vowel blocks (M = 3.71, 95% CIs [3.51, 3.90]) were discriminated better than the variable-vowel blocks (M =2.81, 95% CIs [2.60, 3.03]), β = .90, SE = .11, $t(108) = 8.55$, $p < .001$. Similarly, in the constant-vowel condition, the constant-talker blocks (M = 3.71) were discriminated better than the variable-talker blocks (M = 2.91, 95% CIs [2.70, 3.12]), β = .80, SE = .10, $t(113) = 7.70$, $p < .001$). In addition, the constant-talker-constant-vowel block had significantly higher scores (M = 3.71) than the variable-talker-variable-vowel block (M= 2.67, 95% CIs [2.46, 2.88]), β = 1.04, SE = .11, $t(60) = 9.90$, $p < .001$.

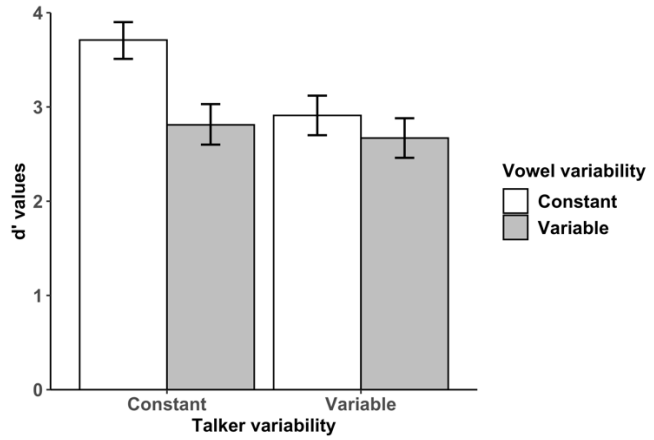


Figure 6.3 Interaction between talker and vowel variability in discrimination by Mandarin listeners. *d'* plotted along the y axis was calculated using a differencing rule for AX tasks (Macmillan & Creelman, 2005). Error bars indicate 95% confidence intervals around the mean.

6.5.3 Discussion

In this experiment, our PAM-based predictions with consideration of phonological overlap types and phonetic overlap scores were upheld. The Two-Category/Non-overlap contrasts (T33-T45 and T33-T21) were better discriminated than Category-Goodness/Non-overlap contrast (T241-T21). Within the same assimilation-overlap category, both overlap scores and our fit index difference scores successfully predicted T33-T45 to be better discriminated than T33-T21.

Native phonological category overlap affects non-native discrimination. Discrimination of T33-T241, a Two-Category/Partial-overlap contrast under high memory load, was worse than those of T33-T45 and T33-T21, Two-Category/Non-overlap, and was as poor as that of T315-T45, UnCategorised-Categorised/Partial-overlap, as expected. In addition, T315-T45 was also discriminated with lower accuracy than T241-T21, Category-Goodness/Non-overlap, as predicted, due to phonological overlap.

Categorisation and discrimination performance were affected by different factors. Memory load affected percent choice in the categorisation task. As we have previously suggested (Chen et al.,

2020), Categorised assimilations reflect a clear *phonological* influence from the native language tone system, whereas UnCategorised assimilations reflect weaker native phonological influence. In addition, the high percent choice and/or goodness rating for a Categorised native response reflect low residual phonetic sensitivity to the difference of the non-native category from the native category. The percent choice of T33 increased for the Categorised native response M55, the only Mandarin level tone, under high memory load, suggesting stronger native phonological constraints. For the Categorised rising Thai tone T45 and falling tone T241, the differences were small between two memory load conditions. There was a sizeable difference in T21 between two memory load conditions (around 14%). T21 was Categorised more consistently under high memory load. The UnCategorised Thai tone T315 showed small differences between two memory load conditions as expected because UnCategorised assimilations reflect weaker native phonological influence.

Talker and vowel variability did not affect assimilation. We reason that as categorisation explicitly requires participants to use their native phonological categories, they were likely to use the phonological mode of perception. Thus, according to ASP (Strange, 2011), they were attending to high level phonological information when making judgements and were less susceptible to low level linguistically irrelevant variability, i.e., talker and/or vowel variability. We further argue that the low level variabilities are accounted for by a similar mechanism as in cross-language categorisation because perceivers can assimilate indexical properties of unfamiliar talkers into the key indexical features of their native speech community (Best, 2015). In other words, for categorisation, Mandarin listeners may rely on mechanisms used in their native language to maintain phonological constancy, i.e., the ability to keep word identity intact, and accommodate to lexically irrelevant variability.

On the other hand, discrimination was unaffected by memory load, differing from prior work with consonants (Werker & Logan, 1985), but consistent with previous findings on tone perception (Lee et al., 1996), supporting the cue-duration hypothesis that the longer the duration of a cue, the less likely it is to decay in short-term memory (Fujisaki & Kawashima, 1970). The acoustic cues of lexical tones, i.e., fundamental frequency, extend over the entire duration of sonorant syllables, i.e., [ma:] and [mi:], should be less likely to decay and therefore more stable relative to consonant cues, e.g., formant transitions or voice onset time.

Both talker and vowel context variability affected discrimination of non-native tones, unlike their lack of significant impact on assimilation. In assimilation tasks, more attention is dedicated to grouping perceptually similar objects into the same category at more of a higher phonological level, whereas in discrimination tasks listeners may attend to phonetic differences. Thus, when there are high variabilities in the stimuli, discrimination accuracy is reduced. This supports our hypothesis that high variability in the stimuli will bias listeners to a phonological mode of perception in which they perceived abstract pitch contours and heights and were less sensitive to specific, concrete F0 properties at the phonetic level.

6.6 Experiment 2: Vietnamese speakers' perception of Thai tones

Experiment 2 was designed to test whether the effects observed in Experiment 1 extend to listeners of another tone language that has a different tone system than both Mandarin and Thai. Vietnamese was selected as it satisfies these requirements. This enabled us to test whether the same five Thai tones (i.e., T241-T21, T33-T21, T315-T45, T33-T241 and T33-T45) would form different assimilation patterns in Vietnamese and in turn lead to different discrimination performance in accordance with PAM-based predictions. For Vietnamese listeners, according to the previous study (Chen et al., 2020): T241-T21 was a Two-Category assimilation contrast; T33-T21 was an Single-Category assimilation contrast; T315-T45 was a UnCategorised-Categorised assimilation contrast; T33-T241 was a Two-Category assimilation contrast ;T33-T45 was a UnCategorised-Categorised assimilation contrast.

6.6.1 Method

6.6.1.1 Participants

We tested 32 native speakers of Southern Vietnamese, divided evenly into two between-subject memory load conditions with 16 participants each (Low memory load: M_{age} = 24.4 years, Age range: 18-44, 12 females, 4 males; High memory load: M_{age} = 27.3 years, Age range: 18-57, 13 females, 3 males). They completed the same discrimination and categorisation tasks as in Experiment 1.

6.6.1.2 Stimuli and Procedure

These were the same as in Experiment 1. The only difference was that for the categorisation task,

the Vietnamese participants were asked to categorise Thai tones using their five[22] native tone categories. The stickers were placed on the keys in the same line on the keyboard, i.e., "f", "g", "h", "j", "k" for the Vietnamese participants.

6.6.2 Results

6.6.2.1 Categorisation

To determine assimilation type of Thai tones, the same criteria were employed as Experiment 1, except that chance level for Vietnamese speakers was 20% as Southern Vietnamese has five tones. Missing data accounts for 0.5% of the total data. We followed the same procedures as in Experiment 1 for determining Categorised or UnCategorised assimilation (see Appendix B, Table B.4, Table B.5, and Table B.6 for details on the statistical tests). Three Thai tones were Categorised into native Vietnamese tone categories under both memory loads: T21 was Categorised to SV22, T241 was Categorised to SV44, and T315 was Categorised to SV214. Both T33 and T45 were UnCategorised under low memory load but Categorised to SV22 and SV21 respectively under high memory load (see Table 6.5).

We fitted the data with multinomial regression models as in Experiment 1. The full model was built with Vietnamese tone choices as a dependent measure, and memory load, talker variability, vowel variability and Thai tones (T21, T33, T45, T241, T315) as fixed factors. Four additional models were built each with one fixed factor subtracted and these models were compared to the full model to determine the effects of cognitive load. Both memory loads, $\chi^2(80) = 190.21, p < .001$, and Thai tone types, $\chi^2(128) = 7243.49, p < .001$, showed significant effects in Likelihood ratio

[22] The standard dialect of Vietnamese (Northern dialect) has six tones but Chen et al (2020) showed that Southern Vietnamese speakers produced both V214 (hỏi) and V415 (ngã) as a tone merger (no significant acoustic differences) consistent with previous reports.

tests. However, the talker variability and vowel variability effects were not significant.

Table 6.5 Assimilation of Thai tones into Vietnamese tone categories under low versus high memory loads. Categories significantly above chance (20%) are in bold; "*" = Categorised tone. Assimilations: C = Categorised, U = UnCategorised. Ratings: 1 = Poor, 7 = perfect; mean ratings are displayed. "-" = no response.

	Thai	T45		T33		T21		T315		T241	
	Response	%	rating	%	rating	%	rating	%	rating	%	rating
Low memory load	SV44	6.3	3.5	**43.5**	5.4	4.3	3.6	0.2	1	**82.7***	5.7
	SV22	2.2	2.4	**51.8**	5.3	**88.7***	5.4	1.8	3.4	13.2	5.4
	SV35	**24.6**	4.7	1.1	2.2	0.5	4.5	7.6	5.8	1.8	3.5
	SV21	**42.7**	4.3	3.6	3.6	6.3	3.9	5.6	3.5	1.6	3.8
	SV214	**24.1**	4.3	-	-	0.2	1	**84.8***	5.2	0.7	3.3
Assimilation		$U_{clustered}$		$U_{clustered}$		C		C		C	
High memory load	SV44	0.4	1	**38.1**	5	3.4	3.1	0.2	1	**81.1***	4.8
	SV22	0.4	2.5	**61.6***	5	**93.7***	4.7	-	-	16.6	4.5
	SV35	15.4	3.9	-	-	-	-	1.6	4.3	1.8	3
	SV21	**60.9***	4	0.2	1	1.6	2.3	6.8	2.8	0.4	1.5
	SV214	22.8	3	-	-	1.3	2.5	**91.4***	4	-	-
Assimilation		C		C		C		C		C	

6.6.2.2 Predictions for discrimination

First, we identified phonological assimilation and overlap types for the Vietnamese group based on perceptual assimilation results. The T33-T45 contrast resulted in UnCategorised-UnCategorised/Non-overlap assimilation under low memory load but Two-Category/Non-overlap under high memory load. The T241-T21 contrast was a Two-Category/Non-overlap assimilation under both memory loads. The T315-T45 contrast yielded UnCategorised-Categorised/Partial-

overlap assimilation under low memory load but Two-Category/Non-overlap assimilation under high memory load. T33-T21 was UnCategorised-Categorised/Partial-overlap under low memory load but Single-Category/Non-overlap under high memory load, where the difference in ratings was not significant, $t(28) = -0.88$, $p = .38$. T33-T241 was UnCategorised-Categorised/Partial-overlap under low memory load but Two-Category/Partial-overlap under high memory load. Thus, the predictions for discrimination accuracy based solely on the assimilation type and phonological overlap are, from good to poor: under low memory load {UnCategorised-UnCategorised/Non-overlap: T33-T45} = {Two-Category/Non-overlap: T241-T21} > {UnCategorised-Categorised/Partial-overlap: T315-T45 = T33-T21 = T33-T241} and under high memory load {Two-Category/Non-overlap: T33-T45 = T241-T21 = T315-T45} > {Two-Category/Partial-overlap: T33-T241} = {Single-Category/Non-overlap: T33-T21}.

As in Experiment 1, we also calculated both the mean and relevant 95% confidence intervals for the overlap scores and fit index difference scores (Appendix B, Table B.8), and again used the confidence intervals to adjust predictions about discrimination due to differences in residual phonetic sensitivities. Under low memory load, there were no significant residual phonetic sensitivity differences between T33-T45 and T241-T21, nor among T315-T45, T33-T241 and T33-T21. Thus, our original predictions on discrimination performance remained unchanged. Under high memory load, the overlap scores but not the fit index difference scores showed differences: T33-T45 ($M_{\text{Overlap\_score}} = .007$) was lower than both T241-T21 ($M_{\text{Overlap\_score}} = .204$) and T315-T45 ($M_{\text{Overlap\_score}} = .312$) but the latter two contrasts did not differ. The predictions on discrimination performance should therefore be changed from being equivalent, i.e., T33-T45 = T241-T21 = T315-T45, as predicted only based on assimilation and overlap types, to: T33-T45 > T33-T21= T315-T45. All other predictions listed earlier remain unchanged.

6.6.2.3 Discrimination

The *d′* scores for each participant were calculated separately for each tone pair as the dependent variable. Trials with time-outs were missing data, which accounted for 1.2% of all the data across all participants in this task. We modelled the data using a Linear Mixed Effects Regression (LMER) model and calculated *p* values in the same way as in Experiment 1(see Table 6.6 for *p*-values and Table B.11 for estimates of the model).

Table 6.6 Main effects and interactions for Thai tone contrast discrimination by Vietnamese listeners. "*" indicates significant.

Cognitive factors	F	df		p
Memory load	0.72	1	30	0.40
Talker variability	30.77	1	30	< .01*
Vowel variability	44.82	1	30	< .01*
Tone contrasts	105.92	4	510	< .01*
Memory load × Talker variability	0.09	1	30	0.77
Memory load × Vowel variability	6.40	1	30	0.02*
Talker variability × Vowel variability	19.79	1	510	< .01*
Memory load × Tone contrasts	2.39	4	510	0.05*
Talker variability × Tone contrasts	0.82	4	510	0.51
Vowel variability × Tone contrasts	2.65	4	510	0.03*
Memory load × Talker variability × Vowel variability	0.19	1	510	0.67
Memory load × Talker variability Tone contrasts	0.30	4	510	0.88
Memory load × Vowel variability × Tone contrasts	0.56	4	510	0.69
Talker variability × Vowel variability × Tone contrasts	0.72	4	510	0.58
Memory load × Talker variability × Vowel variability × Tone contrasts	0.22	4	510	0.93

The main effect of memory load was non-significant. However, there were significant main effects

of talker variability, vowel variability, and Thai tone contrasts. The scores for the constant-talker blocks (M = 3.15, 95% CIs [3.00, 3.30]) were significantly higher, suggesting better discrimination, than those of the variable-talker blocks (M = 2.72, 95% CIs [2.56, 2.87]). Similarly, the scores for the constant-vowel blocks (M = 3.20, 95% CIs [3.06, 3.35]) were significantly higher than those of the variable-vowel blocks (M = 2.67, 95% CIs [2.51, 2.82]).

To further examine Thai tone contrast main effects and test our predictions based on assimilation patterns, multiple comparisons with Tukey adjustments were conducted. All combinations of Thai tone contrasts were significantly different from each other, except for the comparison between T33-T21 and T33-T241 (see Appendix B, Table B.12). The results reflect the following overall pattern: T33-T45 (M = 4.08, 95% CIs [3.86, 4.30]) > T241-T21 (M = 3.64, 95% CIs [3.46, 3.82]) > T315-T45 (M = 2.69, 95% CIs [2.47, 2.92]) > T33-241(M = 1.98, 95% CIs [1.83, 2.13]) = T33-T21 (M = 2.28, 95% CIs [2.05, 2.51]).

In addition to main effects, multiple comparisons were conducted with Tukey adjustments to break down the talker variability × vowel variability interaction (see Figure 6.4, Panel A). In the constant-talker condition, the constant-vowel blocks (M = 3.59, 95% CIs [3.41, 3.78]) were better discriminated than the variable-vowel blocks (M = 2.71, 95% CIs [2.49, 2.93]), β = .88, SE = .11, $t(107) = 7.91$, $p < .001$. Within the constant-vowel blocks, the constant-talker blocks (M = 3.59) were better discriminated than the variable-talker blocks (M = 2.81, 95% CIs [2.60, 3.03], β = .78, SE = .11, $t(113) = 7.07$, $p < .001$). In addition, the constant-talker-constant-vowel block had significantly higher scores (M = 3.59) than the variable-talker-variable-vowel block, (M= 2.62, 95% CIs [2.40, 2.85]), β = .96, SE = .11, $t(60) = 8.67$, $p < .001$.

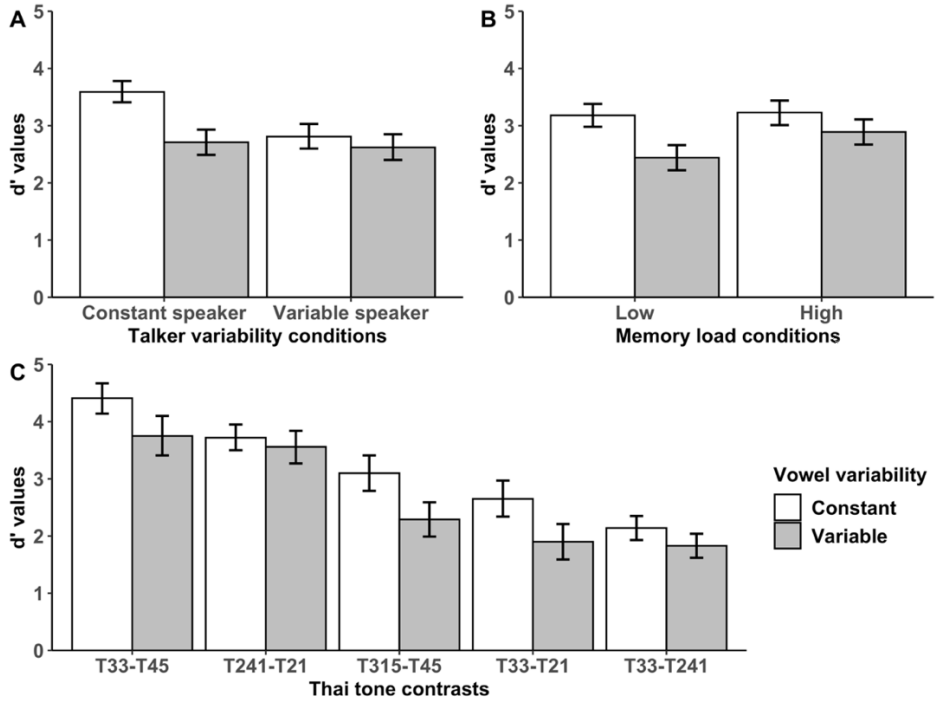


Figure 6.4 Interactions between vowel variability and talker variability (panel A), vowel variability and memory load (panel B), vowel variability and tone contrasts (panel C). *d'* plotted along the y axis was calculated using a differencing rule for AX tasks (Macmillan & Creelman, 2005). Error bars indicate 95% confidence intervals around the mean.

Similarly, we conducted multiple comparisons with Tukey adjustments on the memory load × vowel variability interaction (see Figure 6.4, Panel B). Under low memory load, the constant-vowel conditions (M = 3.18, 95% CIs [2.98, 3.38]) were better discriminated than the variable-vowel conditions *(*M = 2.44, 95% CIs [2.22, 2.66], β = .74, SE = .114, $t(30) = 6.52$, p <.001). Similarly, when the memory load was high, constant-vowel (M = 3.23, 95% CIs [3.01, 3.44]) conditions were also better discriminated than the variable-vowel conditions (M = 2.89, 95%, CIs [2.67, 3.11], β = .33, SE = .11, $t(30) = 2.95$, p =.03). In addition, the constant-vowel blocks under high memory load (M = 3.23) were better discriminated than the variable-vowel blocks under low memory load (M = 2.44), β = .79, SE = .20, $t(41) = 4.02$, p =.001.

We also conducted multiple comparisons with Tukey adjustments to break down the interaction

between vowel variability and tone contrasts (see Figure 6.4, panel C). For the T33-T21 contrast, the constant-vowel blocks (M = 2.65, 95% CIs [2.34, 2.97]) were discriminated better than the variable-vowel blocks (M = 1.90, 95% CIs [1.59, 2.21], β = .75, SE = .18, t(370) = 4.3, p < .01); for T315-T45, the constant-vowel blocks (M = 3.10, 95% CIs [2.79, 3.41]) were discriminated better than the variable-vowel blocks (M = 2.29, 95% CIs [1.99, 2.59], β = .81, SE = .18, t(370) = 4.61, p < .001); and for the T33-T45 contrast, the constant-vowel blocks (M = 4.41, 95% CIs [4.14, 4.67]) were discriminated better than the variable-vowel blocks (M = 3.75, 95% CIs [3.41, 4.10]), β = .65, SE = .18, t(370) = 3.74 , p < .01.

Lastly, we did multiple comparisons with Tukey adjustments to break down the just significant (p = .05) interaction between memory load and tone contrasts. However, there were no significant differences when comparing the same tone between the two memory loads.

6.6.3 Discussion

In Experiment 2, perceptual assimilation patterns were more greatly modulated by memory load than the Mandarin group in Experiment 1. Given that Categorised assimilations reflect stronger native phonological constraints than UnCategorised assimilations, T45 and T33 were UnCategorised under low memory load but became Categorised assimilations under high memory load, reflecting stronger native phonological constraints in the latter case. On the other hand, high percent choice of Categorised native responses reflects low phonetic sensitivity to deviations of non-native tones from native categories. The percent choice of Categorised native response categories for T21 and T315 increases under high memory load, suggesting reduced residual sensitivity to phonetic deviations between native and non-native categories.

PAM-driven predictions based Vietnamese assimilation patterns with consideration of phonological overlap types and phonetic overlap scores were supported by the discrimination

results. The T33-T21 contrast was a Single-Category assimilation contrast under high memory load and was discriminated more poorly than T33-T45, T241-T21 and T315-T45 (all assimilated as Two-Category/Non-overlap under high memory load) as predicted. Phonological overlap reduced discrimination accuracy of non-native contrasts. The T241-T21(Two-Category/Non-overlap under low memory load) contrast was discriminated more accurately than T315-T45 (UnCategorised-Categorised/Partial-overlap under low memory load). Similarly, the T33-T241 contrast was an UnCategorised-Categorised/Partial-overlap assimilation under low memory load but Two-Category/Partial-overlap under high memory load. In both cases, phonological overlap in native response categories reduced its discrimination accuracy to be as poor as T33-T21. Phonetic sensitivity as indicated by overlap scores but not fit-index difference scores, successfully predicted more accurately discrimination of T33-T45 (Two-Category/Non-overlap under high memory load) than that of T241-T21 (Two-Category/Non-overlap under high memory load).

As in Experiment 1, categorisation but not discrimination was affected by memory load manipulations. Manipulating the timing of categorisation responses had an effect because listeners went to phonological-level processing under high memory load, and consequently the results were more constrained by native phonological factors than those under low memory load. In discrimination, we found no main effects of memory load, suggesting that tones may persist longer in working memory than research with consonants indicates.

In addition, talker variability and vowel variability affected discrimination but not categorisation. We speculated that categorisation elicited phonological processing as it required listeners to assimilate non-native phones into native phonological categories and listeners may rely on mechanisms used in their native language to maintain phonological constancy, i.e., the ability to keep word identity intact and accommodate to linguistically irrelevant variability. On the other

hand, discrimination only requires listeners to compare two non-native phones with more focus on the phonetic level details. In variable talker/vowel blocks, listeners tended to use a phonological mode of perception and thus were less sensitive to phonetic details and more constraint by native language. Discrimination in this case was less accurate than in constant talker/vowel blocks where listeners shifted to use a phonetic mode of perception and were more sensitive to the specific, concrete F0 properties.

6.6.4 Cross-language comparison of Thai tone contrast discrimination

The analysis so far has revealed how assimilation patterns modulated discrimination of tone contrasts within each language group. In addition, the discrimination of each tone contrast across two language groups was compared to see whether cross-linguistic PAM-based predictions about differences between the two language groups are also supported. If so, the same tone contrast assimilated differently by the two groups would lead to different discrimination performance.

To this end, we ran a Linear Mixed Effect Regression (LMER) model with *d′* as the dependent variable, and native languages, memory load, talker variability, vowel variability and tone contrasts (T241-T21, T33-T21, T315-T45, T33-T241, T33-T45) as fixed factors, and participant as a random factor including random slopes for within-subject fixed factors, i.e., talker and vowel variability conditions (see Appendix B, Table B.13 and B.14 for full statistical details). There were main effects of talker, $F(1, 60) = 71.07$, $p < .001$, vowel variability, $F(1, 60) = 101.33$, $p < .001$, and Thai tone contrasts, $F(4, 1020) = 198.98$, $p < .001$, and two interactions, talker variability × vowel variability, $F(1, 1020) = 39.80$, $p < .001$, and language × Thai tone contrasts, $F(4, 1020) = 32.80$, $p < .001$, but no main effect of language group, indicating that neither language group had an overall advantage in discriminating all five Thai tone contrasts, but that their performance diverged for at least some tone contrasts.

We conducted multiple comparisons with Tukey adjustments to explore the interaction between language groups and Thai stimuli (see Appendix B, Table B.15 for the full comparisons). The discrimination of two contrasts, T241-T21 and T33-T21, showed significant differences between the two language groups, while discrimination of the other three contrasts T315-T45, T33-T241, and T33-T45 did not differ between groups. Consistent with PAM predictions, the T241-T21 contrast was better discriminated by Vietnamese listeners (assimilated as Two-Category/Non-overlap, M = 3.64, 95% CIs [3.46, 3.82]) than by Mandarin listeners (Category-Goodness/Non-overlap, M = 3.05, 95% CIs [2.85, 3.26]), β = - .59, SE = .17, t(153) = -3.40 , p = .03. Conversely, the T33-T21 contrast was discriminated much better by Mandarin (Two-Category/Non-overlap, M= 3.42, 95% CIs [3.21, 3.63]) than by Vietnamese listeners (UnCategorised-Categorised/Partial-overlap or Single-Category/Non-overlap, M= 2.28, 95% CIs [2.05, 2.51]), β = 1.14, SE = .17, t(153) = 6.58 , p < .01, as predicted.

6.7 General Discussion

The aims of the two experiments in the present study were to examine the native phonological and phonetic influence on the perceptual assimilation and discrimination of non-native lexical tones and how memory load, talker and vowel variability shift listeners between a phonological versus phonetic mode.

6.7.1 Native languages influence on non-native tone perception

Native language tone systems shaped the perception of non-native tones for both language groups as predicted by PAM principles. When two non-native tones were categorised to the same native tone category (Single-Category or Category-Goodness), discrimination was less accurate than when two non-native tones were neatly categorised into two native tone categories with no overlap in response categories (Two-Category).

In addition, perceived phonological overlap, i.e., complete, partial and no overlap, in listeners' assimilations of non-native tone contrasts affects their discrimination. Contrasts with greater phonological overlap in categorisation were discriminated more poorly than those with a smaller degree of overlap within each of the three assimilation types: Two-Category, UnCategorised-Categorised and UnCategorised-UnCategorised. The effect of phonological overlap in reducing discrimination accuracy can be large. For example, T33-T241 was among the worst discriminated contrasts for both groups even though it was assimilated as Two-Category with partial overlap under high memory load.

Phonetic effects as indicated by both overlap scores and fit index difference scores can refine some predictions on discrimination of the same/similar assimilation-overlap type contrast. For example, T33-T45 and T33-T21 are both Two-Category/Non-overlap assimilations for Mandarin listeners. Both overlap scores and fit index difference scores successfully predicted T33-T45 to be better discriminated than T33-T21, suggesting that residual phonetic sensitivity to non-native phones from native categories thus should be considered to make more precise PAM predictions of contrasts of the same assimilation-overlap type.

Comparing discrimination performance across two language groups provides strong evidence for native tone system influences. For the same non-native tone contrast, different assimilation patterns between the two language groups led to different levels of discrimination accuracy, consistent with PAM principles. For example, T21 was assimilated into the high falling tone M51 by Mandarin listeners who have no low level tones but was assimilated to the low level tone SV22 by Vietnamese listeners who have a height contrast between two levels tones. Additionally, mid-level tone T33 was assimilated as high level tone M55 by Mandarin listeners but as low level tone SV22 by Vietnamese listeners under high memory load and split between SV44 and SV22 under

low memory load. Thus, the T33-T21 contrast forms a Two-Category/Non-overlap assimilation for Mandarin listeners but an UnCategorised-Categorised/Partial-overlap or Single-Category/Non-overlap assimilation for Vietnamese listeners. T33-T21 was better discriminated by Mandarin than Vietnamese listeners in line with PAM predictions. This type of comparison could not be assessed in previous studies that compared non-native listeners of non-tone languages with listeners of a single tone language, because tone and non-tone language listeners assimilate non-native tones at different levels of their native phonological systems (Best, 2019).

The fact that discrimination difficulties clearly vary as a function of listeners' first language backgrounds has important implications for second language teaching of tone languages. Teachers should therefore tailor their pedagogy to help students from different tone language backgrounds to solve their special if not unique problems. For example, in a Thai classroom, if there are both Mandarin-native and Vietnamese-native students, when teaching the Thai tone contrast T33-T21, teachers should allocate more resources (e.g., more check-ups/exercises) to Vietnamese students to make sure they perceive the difference.

6.7.2 Memory load effects on non-native perception

Manipulating the temporal course of the judgements in the tasks exerted different effects on categorisation and discrimination. In the categorisation task, after hearing the stimulus, listeners could start processing immediately and even make a decision before the signal to respond. We speculate that the delayed RI in the categorisation task did not only increase memory load but also lengthened the processing time for the target stimulus so that listeners were more likely to engage in higher level phonological processing, i.e., a phonological mode, and based their decisions on perceived abstract pitch contours. Short RI, on the other hand, pushed listeners to base their categorisation choice more on specific, concrete F0 properties, i.e., a phonetic mode, as they had

insufficient processing time to complete phonological assimilation. Future research could aim to interrupt processing and prevent listeners from making their categorisation decision prior to the signal to respond by adding an intervening task, such as counting digits, between the target stimulus and the signal to respond. The reason we did not include an intervening task in the current study was that we needed the categorisation task to be comparable to the discrimination task, which did not use an intervening task.

On the other hand, in the discrimination task, listeners need to hold the first stimulus in memory in order to compare it with the second stimulus and decide whether the two are the same or different. Longer ISI did not affect discrimination, compatible with previous findings on non-native perception of Mandarin and Cantonese tones (Lee et al., 1996). This lends support to the cue-duration hypothesis (Fujisaki & Kawashima, 1970), according to which phonetic details of longer-duration speech segments such as vowels are better retained in memory. In other words, the long ISI may not have been long enough to yield sufficient fading of phonetic details to reduce performance relative to the short ISI and/or maybe the short ISI was not short enough to prevent some fading of phonetic details. Future research can test this hypothesis by using longer ISI.

Originally, we extended ASP principles with the expectation that both assimilation and discrimination performance would vary under different memory loads as listeners shifted between a phonological and phonetic mode of perception. However, our memory load manipulation affected assimilation but not discrimination. Consequently, our two sets of predictions based on assimilation results under low versus high memory load did not align with discrimination performance, which did not differ between the two memory load conditions. For example, for Vietnamese listeners, better discrimination of T241-T21 than T315-T45 was predicted by assimilation types under *low* memory load, i.e., Two-Category/Non-overlap > UnCategorised-

Categorised/Partial-overlap, but assimilation patterns under high memory load predicted equivalent discrimination performance. On the other hand, assimilation patterns under *high* memory load successfully predicted better discrimination of the contrast T315-T45 (Two-Category/Non-overlap) than T33-T241 (Two-Category/Partial-overlap) but assimilation patterns under low memory load predicted equally good discrimination. Nevertheless, when the two sets of assimilation-based predictions were combined, they did account for discrimination variations across different contrasts. Given that assimilation patterns indicated varying degrees of native language phonological and phonetic contributions under high or low memory load, when they successfully predicted discrimination accuracy, they did reveal the amount of native phonological and phonetic contributions to discrimination. Future research should consider the effects of memory load in modulating assimilation to provide more accurate predictions of discrimination.

6.7.3 Talker and phonetic context variability effect on non-native tone perception

While stimulus variability (talker and vowel variability) affected discrimination performance, categorisation was largely immune to its effects. We propose that the differing stimulus variability effects observed in categorisation and discrimination reflect different levels of processing subserving the two tasks, as has been argued in previous studies (Antoniou et al., 2012, 2013). Categorisation tasks require phonological-level judgements, i.e., categorising non-native tones into native categories. Thus, naïve listeners are more likely to use their native language phonological categories as well as mechanisms used to maintain phonological constancy in the native language when processing the non-native stimulus. In other words, perceivers assimilate not only non-native tones to native categories but also assimilate the phonetic properties of unfamiliar talkers to the key indexical features of their native speech community (Best, 2015).

Since the listeners in the present study were all native tone language speakers, their mental representation of native tones should include variations induced by vowel contexts and talkers. Conversely, in discrimination, listeners simply decide whether the tones in the stimulus pairs were the same or not. Thus, listeners base their decisions on low level phonetic information, i.e., F0 properties, and are more likely to be affected by stimulus variations. With that said, listeners generally did discriminate the same contrast better in constant talker and vowel than in variable talker and vowel blocks. This supports our hypothesis that listeners use more of a phonetic mode in constant talker and vowel blocks, where phonetic details were more reliable than those in variable talker and vowels blocks.

There are two additional differences between categorisation and discrimination tasks regarding talker and vowel variability. The first one is that in the present study, the effects of talker and vowel variability in the categorisation task were cross-trial effects because there was only one tone in each trial. In contrast, in the discrimination task the talker and vowel variability in each block was within-trial, i.e., the A and the X had either the same or different talker(s)/vowel(s). In other words, variability was generally higher in the discrimination than in categorisation task in the variable talker and/or vowel blocks. Additionally, in the discrimination task, two perceptual dimensions were involved and interfered with each other, especially in the variable-vowel condition. Listeners were required to attend selectively to the tone dimension while simultaneously ignoring vowel and/or talker differences between the matching tone stimuli. This required more attentional control in high variability conditions and biased listeners to use a phonological mode of perception.

6.8 Conclusion

Discrimination of non-native tone contrasts is significantly influenced by how non-native tones are assimilated into native tone categories. A novel contribution is that both perceived phonological overlap types and phonetic overlap indices can account for variations in discrimination of non-native contrasts of the same assimilation types, suggesting that both phonological and phonetic factors should be considered when predicting non-native discrimination. Moreover, delaying responses in categorisation, i.e., under high memory load, leads listeners to process the stimuli in a phonological mode, resulting in more phonologically-based assimilations, whereas pushing listeners to respond as in the low memory load condition leads them to base their choices more on the phonetic properties of the stimuli. On the other hand, listeners were not affected by talker or vowel variability in categorisation, which suggests that they used their native language phonological categories to accommodate these variations. Unlike categorisation, high stimulus variability in the discrimination task biased listeners to a phonological mode of perception whereas memory load did not have any effect. Although native language influences remain a primary factor as predicted by PAM, memory load and stimulus variability biased categorisation and discrimination differently toward phonological versus phonetic mode, consistent with the principles of ASP. The current findings thus have substantive implications for theories of tone perception as well as for second language lexical tone learning. Native language phonological and phonetic factors as modulated by phonological and phonetic modes in categorisation and discrimination should be considered when researching and teaching non-native tone perception and learning.

Chapter 7. Cognitive factors in Thai-naïve Mandarin and Vietnamese speakers' imitation of Thai lexical tones

Juqiang Chen[a*], Catherine T. Best[a,b*], Mark Antoniou[a]

[a]Western Sydney University, The MARCS Institute for Brain Behaviour and Development

[b]Haskins Laboratories, New Haven CT, USA

Under review: Laboratory Phonology

7.1 Introduction

To learn to speak a second language (L2), language learners not only need to perceive phonemes but also to produce them. The Speech Learning Model (SLM, Flege, 1995) posits that without accurate perceptual "targets", production of non-native sounds will be inaccurate. In practice, L2 learners start by imitating the words produced by native speakers. Imitation reveals a natural link between speech perception and production, thus providing an excellent opportunity to examine that link in non-native speech learning without mediation by orthographic knowledge.

7.2 Native language constraints on speech imitation

How non-native speakers, especially beginning L2 learners, link perception and production is a central, but as yet unresolved, theoretical issue. L2 speech learning theories have made explicit or implicit claims about this link. SLM (Flege, 1995) explicitly claims that the native language influences on non-native production can be traced back partially to how the non-native phone is perceived. If a non-native phone is equivalently classified as either identical or as similar to the acoustically closest native (L1) category (category assimilation), a single phonetic category will be used to process perceptually linked L1 and non-native (L2) phones, also called diaphones in SLM. In this case, a ‘‘merged’’ category will develop over time that subsumes the phonetic properties of the perceptually linked L1 and L2 phones (Flege et al., 2003). On the other hand, if a non-native phoneme is new to the native phonological system, it will be substituted by a range of variants in the early stages of L2 learning. Ultimately, a separate L2 phonetic category will be established for the new phone. In this case, SLM predicts that the newly established L2 category and the nearest L1 speech category deflect from each other in the learner’s phonetic space.

The Perceptual Assimilation Model (PAM; Best, 1995) was originally created to account for native language influences in non-native speech perception by naïve listeners. The more recently developed PAM-L2 (Best & Tyler, 2007) extends PAM principles to account for L2 speech learning, as well as to production due to its direct-realist meta-theoretical assumption that perceivers detect articulatory properties in speech (Best, 1995; Fowler, 1986). PAM considers native language influences on non-native perception at both phonological and phonetic levels. If a given non-native phone is perceived as corresponding to a single native phonological category, it is considered a Categorised assimilation. Yet within that native phonological constraint, listeners will nevertheless display residual sensitivity to within-category phonetic variations from the native phoneme it is assimilated to, commensurate with the magnitude of phonetic discrepancy from "good" native exemplars. In Categorised assimilation, if the non-native phone is perceived as a good exemplar of that native category, then no further perceptual learning will occur for that non-native phone. But if it is perceived as a phonetically deviant exemplar of that native category, some learning will be possible. On the other hand, if a non-native phone is not assimilated cleanly to any single native phonological category (UnCategorised assimilation), the non-native phone will be less susceptible to native language influence and be easier to learn, because UnCategorised assimilations reflect weaker native phonological influence than do Categorised assimilations.

Imitation of synthetic consonant and vowel continua is also constrained by native phonological categories, even in native speakers, who fail to imitate within-category phonetic variations among a continuum's stimulus items. Instead, their imitations reflect their native phonological categories. For example, when English and Spanish monolinguals and Spanish speakers of L2 English imitated a stop consonant voice onset time (VOT) continuum ranging from /da/ to /ta/, their productions did not show a linear incremental increase in VOT, as the stimuli did, but instead

reflected the VOT categories observed in their perception of native voicing Spanish versus English category boundaries (Flege & Eefting, 1988). Similarly, when Finnish children and adults imitated a synthetic Finnish /æ/ to /ɑ/ vowel continuum, they showed categorical imitation patterns that matched their perception of native phonological categories along the continuum (Alivuotila et al., 2007).

Imitation of non-native phonemes is also modulated by participants' native phonological systems. For example, when asked to imitate eight American English vowels /i, ɪ, e, ɛ, æ, ʌ, ɑ, u/, native Mandarin speakers showed the influence of their native language (Jia et al., 2006): /ɛ/ and /æ/ have no Mandarin counterparts and were imitated less accurately than /i/ and /u/, which have counterparts in Mandarin. Moreover, they found a positive correlation between perception and production accuracy, suggesting that non-native phonemes that are difficult to relate perceptually to the L1 are also difficult to imitate.

Despite ample evidence that native phonology constrains non-native imitation, there is also evidence suggesting listeners can bypass native phonological influence and produce phonetically accurate imitations. English listeners were able to imitate English voiceless stops with artificially extended VOTs accurately, as reflected in their lengthened VOTs compared to imitations of originally non-extended stimuli (Fowler et al., 2003) or read speech words (Shockley et al., 2004).

Most imitation studies have examined perception-production relationships for consonants and vowels. Although lexical tones exist in about 70% of languages in the world (Yip, 2002), only a very small number of studies have investigated perception-production relationships in imitation of non-native tones. As with non-native consonants and vowels, non-native perception of lexical tones is constrained by native language phonological and phonetic factors (Reid et al., 2015; So & Best, 2010a, 2010b). However, imitation involves both perception and production, and non-native

tone imitation could bypass some aspects of phonological encoding in identification, as has been found in previous studies with consonants and vowels. It is, therefore, unresolved whether and/or how non-native tone imitation is affected by native phonological and phonetic influences. To investigate this issue, it is desirable to compare imitation of non-native tones with their perceived phonological and phonetic similarity to native tones. In addition, participants should ideally be speakers of other tone languages because lexical tones do not exist in the native phonological systems of non-tone languages which are comprised of only consonants and vowels (Best, 2019).

However, most previous studies on non-native tone production/imitation have been conducted with non-tone language speakers. Hao and de Jong (2016) found that the imitation of Mandarin tones by English speakers was phonetically more accurate than their identification of Mandarin tones. In addition, their imitations were better than their production of Mandarin tones in read speech. The authors argue that this implies imitation can bypass some aspect of native phonological constraints. However, since English is a non-tone language, thus lacking a phonological tier of lexical tones, by definition there are no native phonological biases that could affect performance. Moreover, the study did not have perceptual assimilation results, so altogether the issue of native language phonological and phonetic impact on non-native tone imitation remains unresolved.

In another study, which did examine participants from another tone language, Cantonese native speakers who had learned Mandarin as an L2 identified, imitated and read Mandarin tones (Hao, 2012). The correlation between tone identification and imitation was not significant, and learners were better at imitating than identifying and reading Mandarin tones, suggesting that imitation does indeed bypass some aspects of native phonological constraints. In addition, these Cantonese speakers assimilated Mandarin high level tone into Cantonese high level tone, Mandarin rising

tone into Cantonese low rising tone, Mandarin falling rising tone into Cantonese low falling tone, and Mandarin falling tone into high level tone[23]. As verified by native Mandarin speakers, the imitation of Mandarin falling tone was the best, followed by high level and high rising tones, whereas the Mandarin falling rising tone was imitated poorly. Since the Mandarin falling rising tone was Categorised by the speakers as Cantonese low falling tone, which does not have a final rising, imitation of Mandarin falling rising should be affected by their native language system and thus lack accurate realisation of the final rise, according to PAM principles. The poor imitation of Mandarin falling rising tone supported this prediction and suggested that non-native tone imitation was constrained by the native language tone system. However, since participants were Cantonese speakers with varying degrees of exposure to Mandarin, rather than being Mandarin-naïve, their Mandarin proficiency confounds the relation between assimilation and imitation. Imitation by low proficiency learners and naïve participants may be affected more by their native language than imitation by high proficiency learners is. Thus, these results should be interpreted with caution.

In order to further explore how native language phonological and phonetic factors affect non-native imitation, it is essential to test both assimilation and imitation with native speakers of a tone language who are naïve to the tones in the stimulus language. Imitators who are native tone language speakers can assimilate non-native tones into native tone categories and thus are affected by native tones at both phonological and phonetic level. On the other hand, being naïve to the stimulus language ensures that native language influence is well controlled. In the present study,

[23] The phonological high level tone in Cantonese has two allotone variants for Cantonese speakers in Hong Kong who were tested in the cited study. One variant is high level while the other is high falling. Mandarin falling tone was Categorised as the Cantonese high falling variant of the phonological high level tone.

we examined the phonological and phonetic influences of the native tone systems of Mandarin and Vietnamese, two unrelated tone languages, on their speakers' non-native imitation of the tones of Thai, a third language unrelated to both languages and unfamiliar to both groups. In this way, we can effectively control imitators' experience with the stimulus language that may have affected the results of Hao (2012).

Although SLM can be used to predict imitation based on similarities between native and non-native tones, the model focuses solely on phonetic categories and does not explicate the influence of native language at both phonetic and phonological levels. Thus, we extended core principles of PAM (Best, 1995; Faris et al., 2018), which does address both levels, to make predictions about imitation performance based on perceived similarity between native and non-native tones obtained in Chapter 6 with the same target language and the same listeners as in the present study. We disentangled native phonological versus phonetic contributions to non-native tone categorisation and imitation by considering both the phonological influence reflected in type of assimilation, i.e., Categorised vs. UnCategorised, and the native phonetic influences reflected in relative percentage choice and goodness ratings for a given native tone category (Chen et al., 2020).

Phonological influence is generally strong for Categorised assimilations and is weaker for UnCategorised assimilations. Within UnCategorised assimilations, the phonological influence is moderate for UnCategorised$_{\text{focalised}}$ assimilations, in which the non-native phone is still assimilated as primarily similar to a single native category, but choices of that native phoneme fall below the defined categorisation threshold (Faris et al., 2018). In UnCategorised$_{\text{clustered}}$ assimilation, the general native phonological influence is weak because the non-native phone is assimilated to a small set of native categories, which are below threshold but above chance level and thus none of them have unique or strong influence on assimilation. The native phonological influence is very

weak for the UnCategorised$_{dispersed}$ assimilation, because assimilations of the non-native phone category are spread across many L1 response categories all below chance level (Faris et al., 2018).

Within those phonological constraints, listeners will nevertheless retain some residual sensitivity to within-category *phonetic* deviations of the non-native phones from their native categories. Residual native phonetic sensitivity is determined separately based on percent choice and goodness ratings of the chosen categories in the assimilation task.

7.3 A dynamic view of non-native speech processing for imitation

Imitation performance varies from being strongly constrained by native phonology to accurately reflecting phonetic details of stimuli, thus demanding a theoretically dynamic account of the imitation process. The Automatic Selective Perception model (ASP: Strange, 2011), which primarily accounts for performance variations in online speech perception by adult naïve listeners and L2 learners, can be extended to explain variations in imitation as it involves perception. ASP claims that selective perception routines are used to process native and non-native speech, as activated by the perceiver's detection of task-relevant information. When the attention focus of a task is on phonetic differences that are essential to lexical distinctions, i.e., recognition of phonological distinctiveness (Best, 2015; Best et al., 2009) and detection of phonological structures (phonemes, words, etc.), the activated routines constitute the phonological mode of speech perception, in which phonetic variations within a native phonological category are likely to be overlooked. On the other hand, when the task requires listeners to attend to finer-grained non-contrastive phonetic details, the phonetic mode is activated, allowing detection of phonetic variation within native phonological categories and of non-native deviations from native exemplars (see also Werker & Tees, 1984; Werker & Logan, 1985).

Perception precedes production in the process of imitation. It follows from this that the accuracy of perception should impact imitation. Following this logic and extrapolating principles from ASP, we postulate that there are two modes of imitation that are linked to the two modes of selective perception. The balance between the phonological and phonetic modes can be influenced by cognitive factors such as memory load and talker/vowel variability that can shift the balance of processing between abstract phonological categories and concrete phonetic details.

7.3.1 Memory load in imitation

The availability of phonetic details in short-term memory determines listeners' ability to accurately imitate non-native stimuli. Imitators can only retain the rich array of fine-grained phonetic details in short-term memory for a limited time before they rapidly decay (Baddeley, 2010; Baddeley & Hitch, 1974). The longer the interval between the presentation of the stimuli and the imitation, the more likely it is that memory of the full range of phonetic details will fade. We will refer to the amount of time that listeners must wait before imitating as memory load.

Non-native imitation should be phonetically more accurate under low memory load, i.e., when the delay in imitating is brief and rich phonetic details of the target stimulus remain available, than under high memory load, when the delay is longer and phonetic details decay. Immediate imitation is reported to be more phonetically accurate and can bypass native phonological constraints, relative to imitation delayed by an intervening task. When native speakers of Polish imitated English voiceless aspirated plosives /p, t, k/, which are characterised by long-lag VOT values unlike Polish short-lag /p, t, k/ VOT values, their productions were more English-like in the immediate imitation condition (significant increase in VOTs). When they instead had to read out a digit in the interval between the target item and imitating, which imposes a high memory load, the phonetic accuracy of their imitation was significantly impaired (Rojczyk, Porzuczek & Bergier,

2013). As auditory memory decays more quickly than memory of more abstract "encoded" phonological categories, which are more constrained by participants' native phonological systems, the interpretation is that they must rely on their longer-lasting but phonetically-impoverished phonological memory for delayed imitation.

However, we must note that even under low memory load, participants in some studies have failed to imitate non-native phones/features accurately. For example, imitations of Japanese gemination contrasts by Japanese-naïve native speakers of German, which lacks gemination contrasts, deviated greatly from those of native speakers, and did not differ between low and high memory load conditions (Asano & Braum, 2016). This implies that native phonological constraints on imitation can occur even under low memory load.

In a recent perceptual study directly relevant to the present study, perceptual assimilation/categorisation of Thai tones into native tones by Mandarin listeners were based more on perceived abstract pitch contour and height features under high memory load but more on detection of fine-grained phonetic details, i.e., specific concrete F0 properties under low load (Chen, Best, Antoniou, et al., 2019). However, discrimination of Thai tone contrasts was not affected by memory load. The authors argued that because tones, unlike consonants and vowels, extend over the entire sonorant portion of a syllable and are thereby longer in duration, they are less susceptible to the decay of phonetic details in short-term memory, and thus listeners rely on tone phonetic details even under high memory load in discrimination. However, it remains unresolved whether and how memory load will affect imitation of non-native lexical tone targets.

7.3.2 Talker and vowel context variability

Talker variability caused by physiological and biomechanical differences in speakers' vocal tracts (Nusbaum & Morin, 1992) are reported to affect speech perception. According to principles of

ASP, high talker variability should bias listeners to use a phonological mode of perception and perceive more abstract information from the speech rather than low level phonetic information which is too variable. On the other hand, low talker variability should shift listeners to a phonetic mode of perception because the phonetic level details in the speech are more nearly constant and sufficient for perception and consequently listeners are able to focus their attention on those more reliable phonetic details. Stimuli with high talker variability bias perception toward a native phonological mode of perception and result in lower accuracy than those with lower talker variability, even in native tone identification by Cantonese listeners (Wong & Diehl, 2003), as well as poorer discrimination of non-native Thai tones (Chen, Best, Antoniou, et al., 2019).

As imitation closely links perception and production, we hypothesise that when talker variability is high, non-native imitators will shift toward a phonological mode of perception and will be less sensitive to phonetic details. Consequently, their non-native imitation will be influenced by their native phonological perceptual routines and phonetically less accurate. When talker variability is low, however, imitators will shift toward a phonetic mode of perception and will be more sensitive to phonetic details. As a result, their imitations will be less susceptible to native phonological constraints and thus phonetically more accurate.

Vowel context variability can also affect speech perception. Judging tones in variable vowel contexts reduces discrimination accuracy relative to when the vowel environment of the tones being judged is constant (Chen, Best, Antoniou, et al., 2019). We hypothesise that high vowel variability biases listeners to a phonological mode of perception because the low level phonetic information is variable, pushing listeners to instead rely on more abstract phonological information. On the other hand, we posit that low vowel variability biases listeners to a phonetic mode of perception as the phonetic details are less variable and more reliable.

As with the proposed effect of talker variability on tone imitation, we expect vowel variability to affect imitation similarly because perception provides the external input to imitation. With variable vowel contexts, we posit that non-native imitators will be less sensitive to phonetic details and more affected by native phonological perceptual routines. Imitation in this case will be phonetically less accurate and be biased more toward phonological features of the native language. With constant vowel contexts, on the other hand, we expect non-native imitators to be more sensitive to phonetic details and less affected by native phonological constraints. Consequently, their imitation will be more phonetically accurate. Few studies have examined the effects of talker and vowel context variability on non-native speech imitation, particularly on non-native tone imitation by tone language speakers. The present study was designed to test the above hypotheses.

7.4 Lexical tones in Thai, Mandarin and Vietnamese

Many studies in tone perception and production have used Mandarin tone stimuli. In the present study, we selected Thai tones as the stimuli to examine whether previous non-native imitation findings with Mandarin tones can be extended to another language from a different language family. It is also less likely to be familiar to speakers of other tone languages than Mandarin is. Mandarin as well as Vietnamese listeners were recruited as participants because their native tone systems contain both level and contour tones, like Thai, and yet both systems differ from that of Thai. Thus, it is possible for them to assimilate non-native Thai tones differently into their native tone categories phonologically as reflected in Categorised or UnCategorised types and phonetically as reflected in percent choices and goodness ratings.

Thai, Mandarin and Vietnamese differ in the number and types of tones in their native inventories. We used published Chao values (Chao, 1930) for the tones in each language to estimate a priori phonetic descriptions, in which F0 height at tone onset and offset, and sometimes at an intervening

point in the tone, is referenced by numbers 1-5 ranging from low to high. Here, Thai tones are designated with T, Mandarin tones with M, and Vietnamese tones with V and we make a distinction between phonological features in terms of perceived abstract pitch contours, i.e., level, rising, falling, falling-rising and heights, i.e., high, mid, low, versus specific, concrete F0 properties. Thai, the target language, has three phonological level tones, high-level T45, mid-level T33, low-level T21; and two phonological contour tones, rising T315 and falling T241 (Reid et al., 2015). Mandarin, on the other hand, has four tones: level M55, rising M35, falling-rising M214, and falling M51 (Yip, 2002). The Vietnamese imitators in the present study all spoke the Southern dialect, which has five tones: two phonological level tones, high-level V44 (ngang), low-level V22 (huyền); and three phonological contour tones, rising V35 (sắc), falling V21 (nặng), and falling-rising V214 (Nhan, 1984). V214 instantiates the South Vietnamese merger of two Northern/standard dialect tones, V214 (hỏi), and V415 (ngã) (Brunelle, 2009; Chen et al., 2020).

7.5 The present study

The present study examines how participants' native tone systems influence the imitation of non-native Thai tones by Thai-naïve native Mandarin and native Southern Vietnamese (hereafter Vietnamese) speakers, and how this influence is modulated by memory load and stimulus variability. To evaluate imitation performance, we compared key acoustic measures of the imitations with the target stimuli. The less the imitations deviate from the target stimuli, the more closely the phonetic details of the targets have been imitated.

In order to make predictions for imitation performance in consideration of native language constraints, we extrapolated from the principles of PAM/PAM-L2 (Best, 1995; Best & Tyler, 2007) and disentangled native phonological versus phonetic contributions to non-native tone categorisation and imitation by considering the type of assimilation for phonological influence and

the relative percentage choice and goodness ratings for phonetic influence (Chen et al., 2020). Phonological influences are predicted to be strong for Categorised assimilation, which should result in imitation more like native tones, and weaker for UnCategorised assimilation. Within those phonological constraints, strong residual sensitivity to within-category phonetic variations of the non-native phones from the imitators' native tone categories should be indicated by low percent choice and goodness ratings, whereas weak residual phonetic sensitivity should be reflected in high percent choice and goodness ratings. Strong residual phonetic sensitivity should facilitate more accurate imitation of non-native tones.

Moreover, we predicted that imitations would be more phonetically accurate, i.e., less deviant from the target Thai stimulus details under low memory load when fine phonetic details are available in short term memory and imitators can engage in a phonetic mode of perception according to the principles we have extrapolated from ASP. In contrast, under high memory load, phonetic details of the target item will have faded from short memory and imitators should engage in a phonological mode of perception. Consequently, we propose that imitation will be more constrained by native language phonological influences. For the tones that are assimilated into a native category that deviates phonetically from the target tone, imitations will be deviant as well if the phonological mode is strongly engaged.

In addition, when talker or vowel are variable within a test block, requiring participants to process linguistically irrelevant phonetic differences in parallel with the crucial tone-related phonetic details, imitators will engage in a phonological mode of perception according to ASP principles. Consequently, their imitations should be phonetically less accurate and more deviant from the target stimuli. In contrast, when the talker and vowel within a block are constant, we posit that

imitators should engage in a phonetic mode of perception and their imitation should be more accurate, deviating less from the target stimuli.

7.6 Experiment 1: Imitation of Thai tones by Mandarin speakers

7.6.1 Method

7.6.1.1 Participants

Native speakers of Mandarin (n = 32) participated in the experiment; all had participated in a related study on perception of Thai tones (i.e., the experiments in Chapter 6), which may constitute prior Thai experience of limited nature, before this imitation task. They were divided into two groups for each imitation condition (low memory load: M_{age} = 26.6 years, SD = 7 years; 10 females[24]; high memory load: M_{age} = 26.0 years, SD = 6.9 years; 10 females). Participants completed a background questionnaire before the test. The Mandarin-speaking participants were all born and raised in various regions in China (i.e., Henan, Hunan, Jilin, Jiangsu, Jiangxi) but were educated in Mandarin from early childhood through high school, and they used Mandarin on a daily basis. None of them spoke Cantonese. All reported normal hearing and none had more than two years of formal musical training, as musical training can facilitate tone perception and imitation (Gottfried et al., 2004). The experiments were approved by the Western Sydney University Human Research Ethics Committee (HREC12560) and all participants signed a consent form prior to testing and were compensated for their time (AU$20).

[24] Although the stimuli were female utterances of Thai tones, both male and female participants were native tone language speakers and were required to imitate the perceived form of the lexical tone but not the exact acoustic F0 contour. In addition, from a developmental perspective, male infants can naturally and automatically imitate the global form of their mothers' and fathers' words and phrases, including tones, "as faithfully as possible" (for them) when acquiring their native language, despite the even much larger difference in their F0s and other formant values, relative to their parents.

7.6.1.2 Stimulus materials

Mean F0 trajectories of the tones in each language are presented in Figure 7.1. The tones of the listener languages are presented for comparison. The Thai syllables were recorded from four female Thai speakers (M_{age} = 30.3 years, *SD* = 3.8 years) for a separate study (Burnham et al., 2009), and were used here with permission from the authors. They were recorded in citation form in two syllables (/ma/ and /mi/), in a sound-treated booth at the MARCS Institute for Brain Behaviour and Development, Western Sydney University, using a Lavalier AKG C417 PP microphone with the sampling rate of 48 kHz and 16 bit resolution. For acoustic analysis, five tokens per target tone per syllable (/ma:, mi:/) were selected from two participants, four tokens each from the third participant, and six each from the fourth participant, yielding 20 tokens per target tone per syllable (20 tokens × 5 tones ×2 syllables = 200 in total).

In addition, we also recorded productions of each native tone in the syllables /ma/ and /mi/ by four female native speakers of each imitator group's language (Mandarin, M_{age} = 27.0 years, *SD* =2.2 years; Vietnamese, M_{age} = 21.0 years, *SD* =3.0 years) with each of their native tones, eight times each by each speaker, in random order. Mandarin items were elicited via Pinyin, Vietnamese items via the orthography of their language. Vietnamese speakers were asked to read all six Vietnamese tones but were specifically instructed to read them in their southern accent. Thus, there were 64 tokens (2 syllables /ma/ and /mi/ × 4 tones each × 8 repetitions) for each Mandarin informant and 96 tokens (2 syllables /ma/ and /mi/ × 6 tones each × 8 repetitions) for each Vietnamese informant. Mandarin and Vietnamese productions were recorded using a Zoom H4n digital speech recorder with a sampling rate of 44.1 kHz and 16-bit resolution in a quiet testing booth at The MARCS Institute, Western Sydney University. All target syllables were meaningful morphemes in the respective languages.

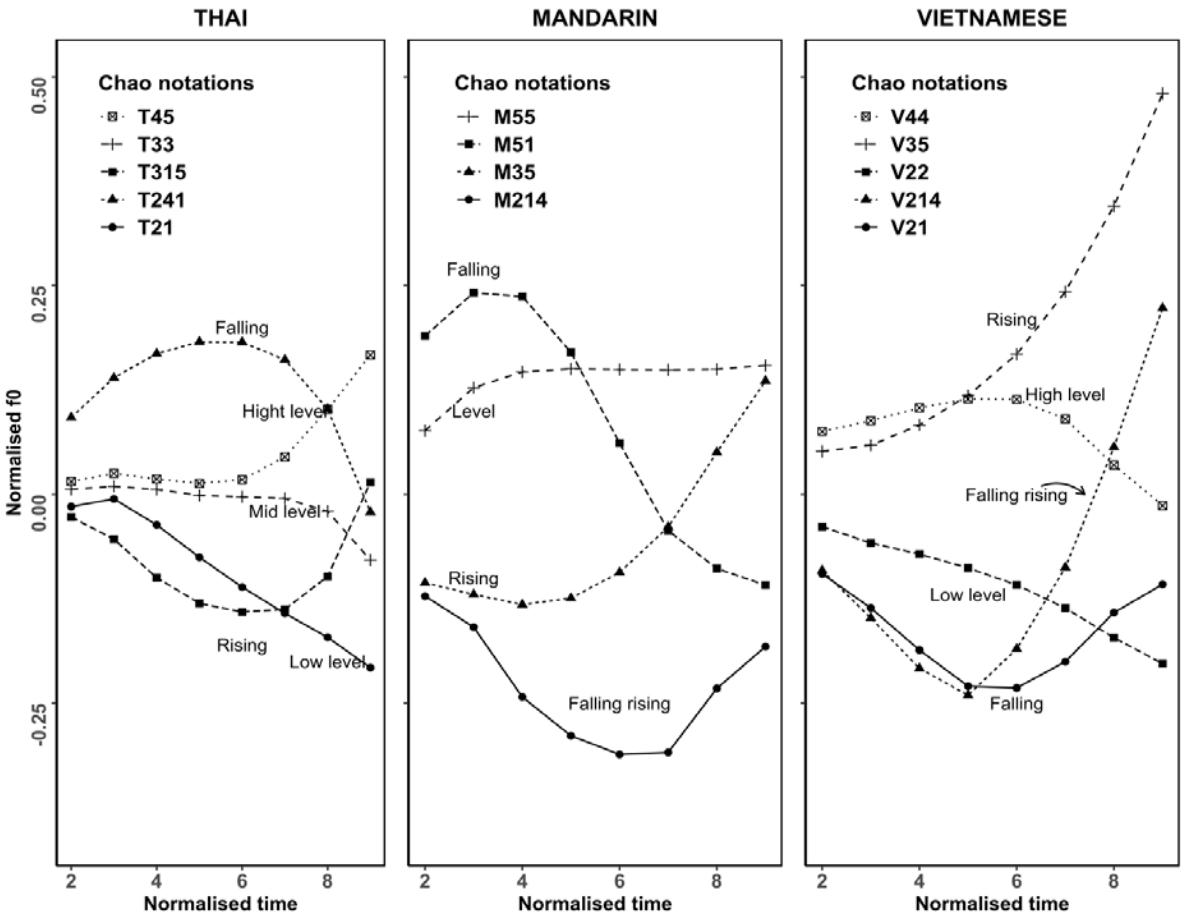


Figure 7.1 Time- and Lobanov-normalised (Lobanov, 1971) F0 contours of Thai, Mandarin and Southern Vietnamese tones[25].

Given that average F0, direction, length, extreme point and slope are reported to be the primary factors affecting the perception of lexical tones (Gandour, 1978), we selected four acoustic measures to characterise tones and their imitations for the present study: duration, F0mean, F0 maximum to minimum excursion (F0excursion), F0 maximum location (F0maxloc) (see Appendix C, Table C.1). F0excursion distinguishes level tones from contour tones, and contour tones such as T241 and T315 can be differentiated by F0maxloc. In a PCA analysis of lexical tones (Chen et

[25] In Chen et al., (2020), we had asked the Southern Vietnamese speakers to produce both V214 (hỏi) and V415 (ngã), but consistent with the reports of merger they showed no significant acoustic differences, so here they were averaged and labelled as the single South Vietnamese phonologically falling rising tone V214.

al., 2018), these acoustic measures outweighed other measures in differentiating Thai, Mandarin, Southern and Northern Vietnamese tones. To make more concrete predictions about the direction of deviations in imitation, we calculated the 95% confidence intervals to compare Thai tones with Mandarin and Vietnamese tones in terms of the four acoustic measures (see Figure 7.2).

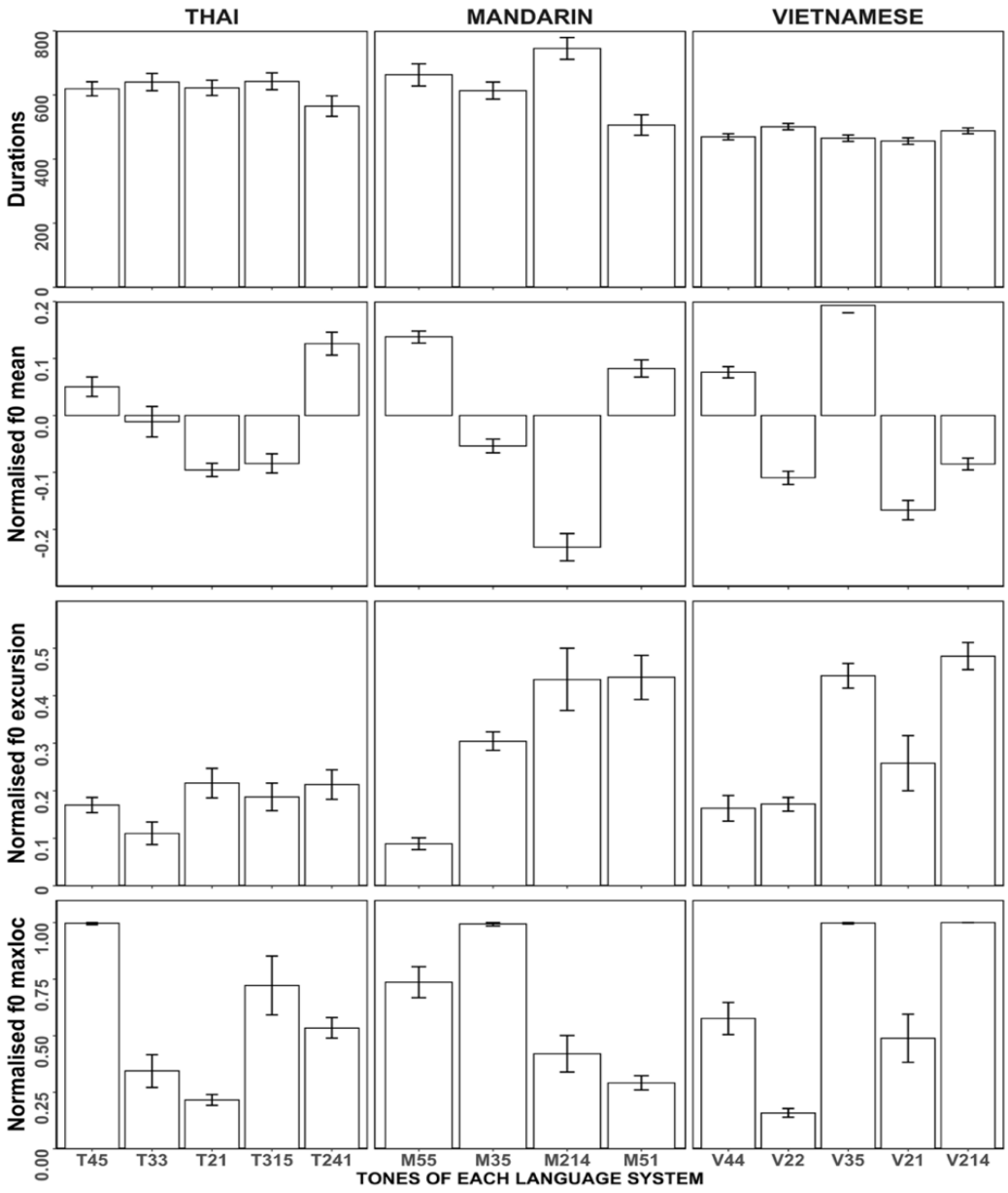


Figure 7.2 Acoustic measures of tones in Thai (20 tokens per tone), Mandarin and Vietnamese (32 tokens per tone). $F0_{mean}$, $F0_{excursion,}$ are Lobanov-normalised Hz values. The error bars indicate the 95% confidence intervals.

7.6.1.3 Predictions

Extending PAM principles to imitation, we argue that if a non-native tone is Categorised as a native tone, then the imitation of that tone will be like the native tone. In Chapter 6, the same Mandarin participants as in the present study perceptually assimilated the same Thai tone stimuli as used in the present study into their five native tone categories under low and high memory loads (see Table 7.1). We used those perceptual data as the basis for predictions about native phonological and phonetic influences on imitations of Thai tones. To quantify residual phonetic sensitivity from the prior perceptual data, we first divided percent choice of the native tones above chance into three ranges: Low, Medium and High. These ranges respectively reflect strong, moderate, and weak residual phonetic effects. The percent choice ranges necessarily differ for the two groups. For Mandarin listeners, Low spanned 25%-49% of choices, Medium 50%-75%, and High 76%-100%. For the category-goodness ratings, we also divided the scale in to three ranges, which apply to both listener groups: Low = 1-2.9, Medium = 3-4.9, and High = 5-7. These ranges reflect strong, moderate and weak residual phonetic effects, respectively.

Under both memory loads, T33 was Categorised as M55; T45 was Categorised as M35; and T241 was Categorised as M51 (Table 7.1). Both percent choice and goodness ratings for the three assimilations were in the high range, suggesting low residual phonetic sensitivity to differences between native and non-native tones. T21 was Categorised as M51 under both memory loads but with percent choice and ratings in the medium range, suggesting moderate residual phonetic sensitivity to differences between native and non-native tones. T315 was an $\text{UnCategorised}_{\text{clustered}}$ assimilation under both memory loads and was split between M35 and M214. For $\text{UnCategorised}_{\text{clustered}}$ assimilations, we predicted the native phonological influence would be relatively weak. Percent choice of native response categories were in low and/or medium ranges,

whereas goodness ratings were in the high range, suggesting moderate residual phonetic sensitivity to differences between native and non-native tones, which is expected to moderately facilitate accurate imitation of non-native tones.

Table 7.1 Assimilation of Thai tones into Mandarin tone categories under low versus high memory loads (from Chapter 6). Categories in bold are choices that were significantly above chance: 25% for Mandarin; "*" = Categorised tone. Assimilations: C = Categorised, U = UnCategorised. Rating: 1 = poor, 7 = perfect; mean ratings are displayed. "-" = no response.

Thai stimulus		T45		T33		T21		T315		T241	
	Response	%	rating	%	rating	%	rating	%	rating	%	rating
Low Memory	M55	0.4	4	**77.3***	5.3	19.9	3.3	-	-	21.6	4.2
	M35	**88.8***	5.8	2.8	2.8	1.6	2.1	**48.6**	5.5	2.9	3.7
	M214	10.3	4.3	0.7	3.3	25.8	3.7	**51.2**	5.4	0.4	3.5
	M51	0.4	5.5	19.2	4.9	**52.7***	4.6	0.2	7	**75.1***	5.6
Assimilation		C		C		C		$U_{clustered}$		C	
High Memory	M55	-	-	**84.7***	5.2	26.4	3.2	0.2	7	**28.1**	5
	M35	**85***	5.3	0.2	5	1.1	3.3	**44.2**	5.1	0.2	2
	M214	14.7	4.8	1.7	5.9	6.3	3.8	**55.6**	5.5	0.5	6
	M51	0.2	2	13.4	4.4	**66.2***	4	-	-	**71.2***	5.1
Assimilation		C		C		C		$U_{clustered}$		C	

For the stimulus variability factors, we systematically manipulated the talker and vowel variability of the Thai target stimuli (constant versus variable blocks). Figure 7.3 shows the talker and vowel variability in the Thai stimuli in terms of their F0 contours.

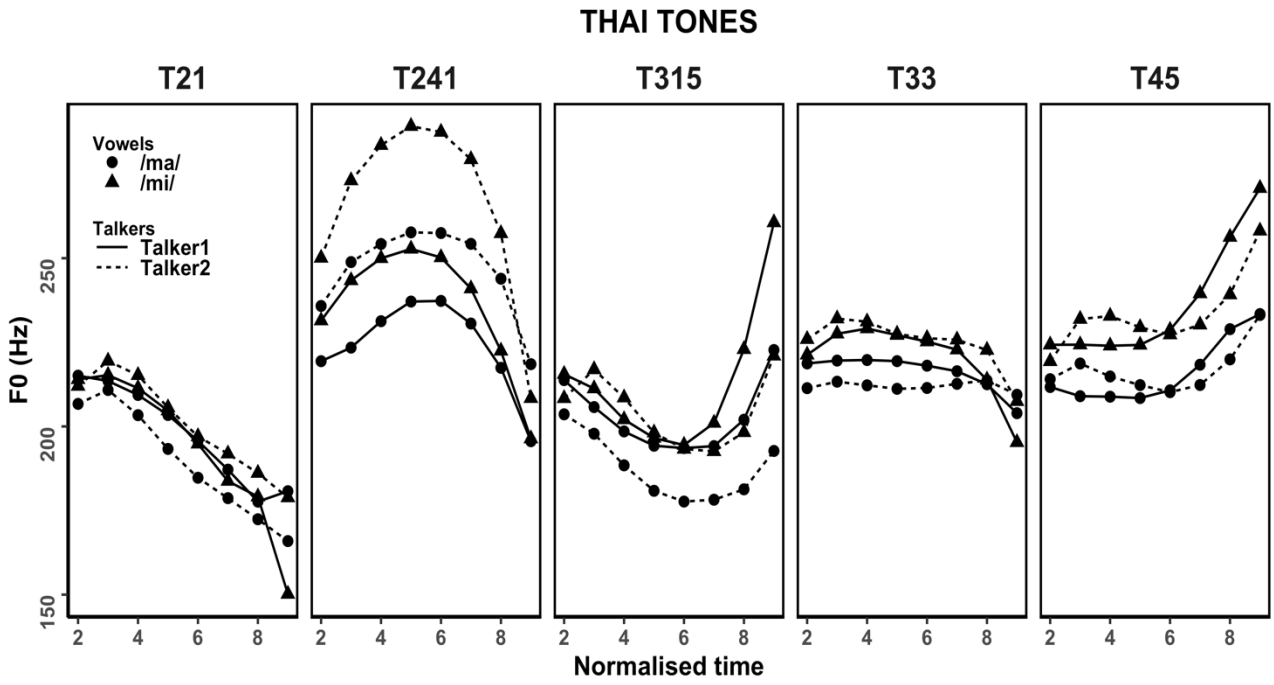


Figure 7.3 Talker and vowel variability in mean F0 contours of Thai stimuli (time-normalised: 5 tokens × 2 talkers × 2 syllables × 5 tones).

7.6.1.4 Procedure

Memory load in the current imitation study was operationalised as the time between the end of the stimulus and a signal for participants to produce their imitation (imitation interval). Under low memory load, a message "Imitate now!" was shown 500 ms after the offset of the stimulus to alert participants to start imitating. Under high memory load, the same message was shown 2000 ms after the offset of the stimulus. Under both memory loads, participants had 3 seconds (timeout) to imitate, and the inter-trial interval was 1 second. We also blocked talker variability (constant = one talker vs. variable = two talkers) and vowel variability (constant = /ma/or /mi/ vs. variable vowels = /ma/ and /mi/) across the experiment.

Participants were instructed to imitate stimuli as faithfully as possible after they heard the auditory stimulus and saw the starting signal. Before the test session, participants completed 10 practice

trials. Then each participant completed 160 imitation trials (2 syllables × 5 tones × 8 conditions × 2 repeats) in total.

Participants were tested individually in testing booths at Western Sydney University, University of New South Wales, and Macquarie University. Stimuli were presented on a Dell Latitude 7280 laptop running E-Prime Professional 2 via Sennheiser HD 280 Pro headphones at 72 dB SPL. Participants' responses were recorded with a portable digital speech recorder (ZOOM H4n) with a 44.1 kHz sampling rate and 16-bit resolution.

7.6.1.5 Acoustic processing

ProsodyPro (Xu, 2013), a *Praat* (Boersma, 2001) script, was used to first extract the F0 contours of the Thai stimuli and their imitations at 10 time-normalised points of F0. In order to make F0 values comparable across different speakers, all F0 values were Lobanov-normalised (Lobanov, 1971), which reflects how much the mean F0 for a given data point varies from the F0 mean of the speaker. According to a previous multidimensional scaling study, duration, F0 mean, F0 direction, F0 extreme endpoint, and F0 slope have been found to correlate with perception of tones by native listeners of Thai and Yoruba (Gandour, 1978) so we used all of these measures. We calculated three F0-based acoustic scores between 10% to 90% of the syllable length from the ten F0 points, the most stable portion, for statistical analyses: $F0_{mean}$, F0 maximum to minimum excursion ($F0_{excursion}$), F0 maximum location ($F0_{maxloc}$).

To quantify the acoustic deviations of the imitations from their target Thai stimuli for use in statistical analyses, we calculated deviation scores following the procedure in Wang, Jongman, and Sereno (2003) by subtracting values of duration and the three Lobanov-normalised F0-related acoustic measures for an imitation token from values for its target stimulus token.

7.6.2 Results

The signed deviation scores for duration, $F0_{excursion}$, $F0_{mean}$, and $F0_{maxloc}$ were each selected as a dependent variable and fitted with a separate linear mixed-effects model. Memory load (low vs. high), talker variability (constant vs. variable), vowel variability (constant vs. variable) and Thai tone (T1-T5) were used as fixed factors. We first ran the analysis with participants as a random factor including random slopes for all within subject factors, i.e., talker and vowel variability (as suggested by Barr et al., 2013). The models converged but were too complex to estimate *p* values for the fixed effects of interest. Thus, we dropped the random slopes and participants was specified as a random intercept in each model. Four models were built to test all main effects and interactions. To calculate the *p*-values for the fixed effects, we again used the Kenward-Roger degrees of freedom approximation (Halekoh and Hojsgaard, 2014), and the *Anova* function from the *car* package in R, with test specified as "F". First, there was a just significant main effect of memory load for $F0_{mean}$ deviation scores, and significant main effects of talker variability for $F0_{maxloc}$ deviation scores and Thai tone contrast for all four deviation scores (see Table 7.2 and for statistical details see Appendix C, Table C.2-C.5). However, the main effect of vowel variability was non-significant for all deviation scores.

The main effect of memory load in terms of $F0_{mean}$ deviation scores showed an unexpected pattern of overall more accurate imitation under high (M = -0.001, 95% CIs [-0.004, 0.001]) than low memory load (M = -0.008, 95% CIs [-0.011, -0.005]) blocks, which should be interpreted in light of the significant memory load × tone type interaction. Similarly, the main effect of talker variability in terms of $F0_{maxloc}$ deviation scores showed a pattern of overall more accurate imitation in variable (M = -0.003, 95% CIs [-0.013, 0.007]) than constant talker (M = -0.065, 95% CIs [-0.072, -0.058]) blocks, which should be interpreted in light of the significant talker variability ×

tone type interaction.

To further examine the tone type main effects, we ran pairwise multiple comparisons with Tukey adjustments for differences in deviation scores among tone types (see Appendix C, Table C.6). All pairwise comparisons among the tones for duration deviation scores were significant. T33 had the largest deviation scores for syllable duration (*M* = .064, 95% CIs [.057, .071]), thus being the least accurate. Syllable duration deviation scores for T241 (*M* = .058, 95% CIs [.051, .065]) were the second largest. Both T21 (*M* = -.014, 95% CIs [-.021, -.006]) and T315 (*M* = -.023, 95% CIs [-.030, -.015]) had negative deviation scores, indicating that the imitations were shorter than the target stimuli. The duration of T45 imitations (*M* = .007, 95% CIs [.0004, 0.015]) was the most accurate, as indicated by the smallest absolute value of deviation scores.

All pairwise comparisons among the tones for $F0_{mean}$ deviation scores were significant. $F0_{mean}$ deviation scores were positive only for T315 (*M* = .046, 95% CIs [.042, .050]), indicating higher $F0_{mean}$ in the imitations than the target stimuli. $F0_{mean}$ deviations were negative for T241 (*M* = -.037, 95% CIs [-.042, 0.032]), T45 (*M* = -.021, 95% CIs [-.025, -.018]), T33 (*M* = -.011, 95% CIs [-.015, -.008]), indicating lower $F0_{mean}$ for imitations than the stimuli. The scores for T21 (*M* = -.0004, 95% CIs [-.004, 0.03]) were very small, indicating the greatest imitation accuracy.

All pairwise comparisons among the tones for $F0_{excursion}$ deviation scores were significant except for the T33-T241 comparison. $F0_{excursion}$ deviation scores were positive for all tones, indicating that Mandarin imitators generally enlarged the range of the F0 contour in the imitations. These scores were largest for T315 (*M* = .171, 95% CIs [.161, 0.181]), followed by T33 (*M* = .114, 95% CIs [.106, .123]), T241 (*M* = .102, 95% CIs [.089, .114]), T45 (*M* = .079, 95% CIs [.069, .088]), suggesting T315 was the least accurately imitated. T21 (*M* = .053, 95% CIs [.045, .061]) was the most accurately imitated, with the smallest $F0_{excursion}$ scores.

All pairwise comparisons among the tones for $F0_{maxloc}$ deviation scores were significant. The $F0_{maxloc}$ deviation scores were positive for T21 (M = .021, 95% CIs [.012, .031]) and T315 (M = .063, 95% CIs [.044, .082]), suggesting that the F0 peak was delayed in imitation than that in the Thai target stimuli. Imitation of T315 was less accurate than T21. In contrast, T241 (M = -.167, 95% CIs [-.179, -.156]), T45 (M = -.068, 95% CIs [-.075, -.061]), T33 (M = -.019, 95% CIs [-.034, -.003]) had negative $F0_{maxloc}$ values, indicating F0 peaks were realised earlier in imitation than in the target stimuli. T241 scores were largest in absolute value, indicating it received the least accurate $F0_{maxloc.}$imitations. T45 was imitated more accurately than T33 on this measure.

Table 7.2 Model details of acoustic measure deviation scores of Mandarin imitators. Significant effects are in bold; marginal effects are in parentheses.

	Duration			$F0_{mean}$			$F0_{excursion}$			$F0_{maxloc}$		
	F	*df*	*p*	*F*	*df*	*p*	*F*	*df*	*p*	*F*	*df*	*p*
Memory (Mem)	0.0	1 ,30	0.989	4.1	1, 30	**0.052**	3.4	1, 30	(0.076)	0.3	1, 30	0.600
Talker (Tlk)	1.0	1, 5026	0.328	0.3	1, 5026	0.571	0.9	1, 5026	0.346	118.6	1, 5026	**<.001**
Vowel (V)	1.0	1, 5026	0.306	1.5	1, 5026	0.218	0.0	1, 5026	0.829	0.3	1, 5026	0.578
Tone	121.9	4, 5026	**<.001**	253.1	4, 5026	**<.001**	97.7	4, 5026	**<.001**	194.6	4, 5026	**<.001**
Mem×Tlk	0.1	1, 5026	0.787	0.1	1, 5026	0.705	0.5	1, 5026	0.471	0.3	1, 5026	0.609
Mem×V	0.6	1, 5026	0.444	0.0	1, 5026	0.941	1.0	1, 5026	0.318	0.1	1, 5026	0.716
Tlk×V	0.6	1, 5026	0.436	0.3	1, 5026	0.565	2.6	1, 5026	0.106	0.1	1, 5026	0.754
Mem×Tone	1.6	4, 5026	0.168	13.8	4, 5026	**<.001**	8.1	4, 5026	**<.001**	2.8	4, 5026	**0.026**
Tlk×Tone	11.7	4, 5026	**<.001**	9.9	4, 5026	**<.001**	1.3	4, 5026	0.250	50.8	4, 5026	**<.001**
V×Tone	0.6	4, 5026	0.649	0.7	4, 5026	0.584	1.0	4, 5026	0.390	0.5	4, 5026	0.754
Mem×Tlk×V	0.7	1, 5026	0.388	0.1	1, 5026	0.745	0.5	1, 5026	0.479	2.6	1, 5026	0.108
Mem×Tlk×Tone	0.1	4, 5026	0.978	0.2	4, 5026	0.923	1.3	4, 5026	0.270	0.4	4, 5026	0.776
Mem×V×Tone	0.2	4, 5026	0.951	0.4	4, 5026	0.815	0.1	4, 5026	0.970	0.2	4, 5026	0.916
Tlk×V×Tone	0.2	4, 5026	0.959	0.3	4, 5026	0.878	0.9	4, 5026	0.468	0.5	4, 5026	0.717
Mem×Tlk×V×Tone	1.0	4, 5026	0.426	0.1	4, 5026	0.970	0.6	4, 5026	0.631	0.5	4, 5026	0.720

To further examine memory load × tone type interactions for $F0_{mean}$, $F0_{excursion}$ and $F0_{maxloc}$ deviation scores, we conducted pairwise comparisons with Tukey adjustments for each tone across memory load conditions for each deviation score (for statistical details see Appendix C, Table C.7). The crucial comparisons are those for the same tone between memory loads with significant differences as described below. $F0_{mean}$ deviation scores of T315 indicated more accurate imitation under low (*M* = .036, 95% CIs [.030, .041]) than high memory load (*M* = .056, 95% CIs [.051, .062]). However, $F0_{mean}$ deviation scores of T45 indicated less accurate imitation under low (*M* = -.033, 95% CIs [-.038, -.028]) than high memory load (*M* = -.010, 95% CIs [-.014, -.005]). Although there were some variations in $F0_{excursion}$ and $F0_{maxloc}$ deviation scores between the two memory load conditions, none of these differences were significant for the same tone type.

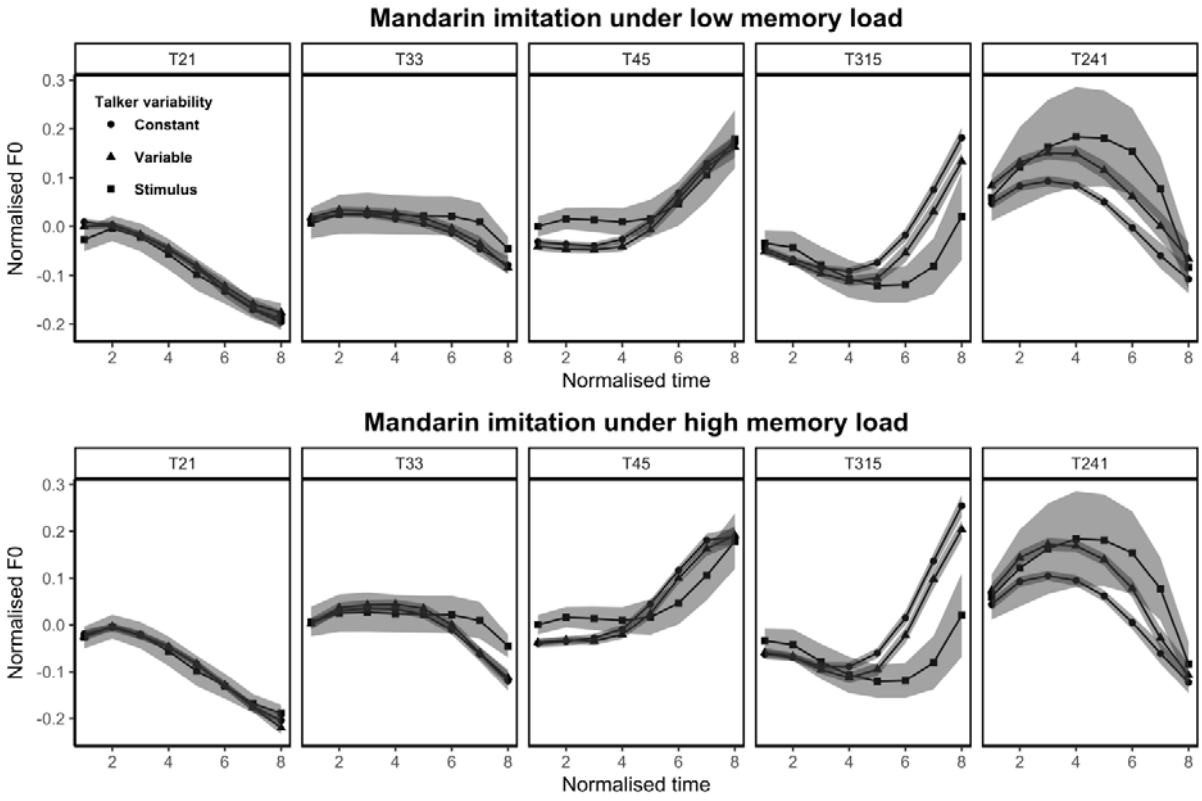


Figure 7.4 The time-and-Lobanov-normalised mean F0 contours of the Thai stimulus tones and their imitations produced by Mandarin participants. Ribbons indicate 95% confidence intervals.

Similarly, to further investigated the talker variability × tone type interactions for duration, $F0_{mean}$, and $F0_{maxloc}$ deviation scores, we ran multiple comparisons with Tukey adjustments to tease the interactions apart (see Figure 7.4 for the general F0 contours, and for full statistical results see Appendix C, Table C.8). Here, the crucial comparisons are those for the same tone between the two talker variability conditions with significant differences as described below. The duration imitation for T21 and T45 were more accurate imitation in constant talker (M_{T21} = .008, 95% CIs [-.002, .018]; M_{T45} = -.005, 95% CIs [-.015, .005]) than variable talker blocks (M_{T21} = -.035, 95% CIs [-.046, -.025]; M_{T45} = .020, 95% CIs [.009, .030]). $F0_{mean}$ deviation score of T241 also indicated more accurate imitation in constant talker (*M* = -.026, 95% CIs [-.031, -.020]) than variable talker blocks (*M* = -.048, 95% CIs [-.056, -.040]).

Imitation of T315 in terms of $F0_{maxloc}$ was more accurate in constant talker (*M* = -.029, 95% CIs [-.038, -.019]) than variable talker blocks (*M* = .155, *SD* = 95% CIs [.119, .190]). However, for T33 and T241, $F0_{maxloc}$ deviation scores indicate *less* accurate imitation in constant talker blocks (M_{T33} = -.081, 95% CIs [-.102, -.061]; M_{T241} = -.188, 95% CIs [-.204, -.172]) than in variable talker blocks (M_{T33} = .044, 95% CIs [.022, .066]; M_{T241} = -.147, 95% CIs [-.162, -.131]).

7.6.3 Discussion

First, memory load showed main effects for $F0_{mean}$ and interacted with tone types for $F0_{mean}$, $F0_{excursion}$ and $F0_{maxloc}$. However, none of variations in $F0_{excursion}$ and $F0_{maxloc}$ deviation scores between the two memory load conditions were significant for the same tone type. Only T315 had smaller deviations and lower $F0_{mean}$ values under low than high memory load, as we had predicted. This suggests that listeners showed high phonetic sensitivity to the target Thai stimuli under low memory load when phonetic details should be available in short memory, which we expected to bias them to use a phonetic mode of perception and imitation. On the other hand, we had expected

listeners to use a phonological mode of perception under high memory load when the phonetic details have faded, and imitation should become less accurate. Because T315 is an UnCategorised$_{clustered}$ tone assimilation with weak native phonological influences, this deviation could not be attributed simply to native language phonological factors, but rather should reflect sensitivities to non-contrastive phonetic details. We speculate that the higher F0 value of a rising tone that appears toward the end of the tone was retained in memory at the time of imitation, by which time the lower F0 onset of the tone had faded under high memory load. Thus, we posit that under high memory load imitators tend to start their imitations of rising tones at a higher F0 and this resulted in an overall higher $F0_{mean}$ relative to the low memory load condition. This hypothesis could also explain why the $F0_{mean}$ of T45 imitation was unexpectedly more accurate but also higher in F0 value under high than low memory load. If listeners had activated a phonological mode of perception, T45 should instead have been affected by the native tone it was assimilated to, i.e., M35 which is *lower* than T45 in $F0_{mean}$. But they imitated with a higher rather than a lower F0. Thus, higher $F0_{mean}$ in imitation of T45 under high memory load cannot be attributed to native language phonological influence but instead to a more phonetic-level tendency to start at a higher F0 when the phonetic details of the lower F0 at the onset of the tone have faded away.

Second, imitation was phonetically more accurate and less susceptible to native phonological influence in constant than variable talker blocks. Extending principles of ASP to imitation, listeners will use a phonetic mode in constant talker blocks because phonetic information is constant and reliable and thus specific, concrete temporal and F0 properties that are better detected. Consequently, imitations should be more accurate in constant than variable talker blocks. Syllable duration for T21 and T45, $F0_{mean}$ for T241, and $F0_{maxloc}$ of T315 in imitation were more accurate in constant than in variable talker blocks as expected. On the other hand, listeners should use a

phonological mode of perception in variable talker blocks in which phonetic information is variable and unreliable and thus listeners have to use their native phonological perceptual routines based on perceptual abstractions of tone contours and heights. Consequently, they will be more constrained by native language phonology. Lower $F0_{mean}$ for T241 in the variable talker block reflected increased native phonological influence because T241 was Categorised as M51 by the same participants, which has a lower $F0_{mean}$ than T241.

Third, despite of manipulation of memory load and stimulus variability, imitation of Categorised tones with high percent choice and goodness ratings reflected strong native phonological influence and low sensitivity to phonetic differences between native and non-native phones. T241 was Categorised to M51 and T45 was Categorised to M35 (Table 7.1) with percent choice and goodness ratings in the high range. Imitations of T241 showed larger $F0_{excursion}$ and earlier $F0_{maxloc}$ than the target stimuli, consistent with the characteristics of M51 relative to T241. Similarly, T45 was imitated with lower $F0_{mean}$ and larger $F0_{excursion}$ than the original stimuli, consistent with the characteristics of M35, to which it had been assimilated. These deviations indicated that imitations of both tones by Mandarin participants were constrained by L1 phonological features.

Conversely, imitation of non-native tones with moderate or low percent choice and goodness ratings was phonetically accurate, suggesting high sensitivity to phonetic differences between native and non-native phones. T21 was Categorised into M51 but with medium range of percent choice and goodness rating. T21 was imitated accurately with low deviation scores and was unaffected by the native category it was assimilated to M51 which had a much larger $F0_{excursion}$ and later $F0_{maxloc}$ than the stimulus. Similarly, T33 was perceptually Categorised to M55 but with moderate goodness ratings. $F0_{mean}$ of imitation of T33 was lower than the target, supporting our

hypothesis that residual sensitivity facilitates accurate imitation, and reduces native phonological impact from M55 which has a higher $F0_{mean}$.

7.7 Experiment 2: Imitation of Thai tones by Vietnamese speakers

Vietnamese differs from Mandarin in their native tone systems, and Vietnamese listeners, accordingly, had assimilated the same Thai tones into their native tone categories differently than Mandarin listeners as indicated in Chapter 6. In the next experiment, Vietnamese participants were asked to imitate Thai tones with the same manipulation of memory load and talker/vowel variability. Their imitation will be analysed with reference to their assimilation patterns to test our PAM and ASP driven hypothesis.

7.7.1 Method

7.7.1.1 Participants

Native speakers of Southern Vietnamese (n = 32) participated in Experiment 2, and were divided into two groups for each imitation condition (low memory load: M_{age} = 24.4 yrs, *SD* = 7.7 yrs; 13 females; high memory load: M_{age} = 27.2 yrs, *SD* = 12.8; 12 females). All had participated in a related study on perception of Thai tones (i.e., the experiments in Chapter 6), which may constitute prior Thai experience of limited nature, before this imitation task. They completed a background questionnaire before the test. All self-reported to have normal hearing and none had experience with Thai or more than two years of formal musical training. Stimulus materials, procedure and data analysis are the same as Experiment 1.

7.7.1.2 Predictions

In the experiment in Chapter 6, the same participants had perceptually assimilated the same Thai tone stimuli used here (see Table 7.3). We used those perceptual data as the basis for predictions about native phonological and phonetic influences on imitations of Thai tones. To quantify residual

phonetic sensitivity, for Vietnamese listeners, we first divided percent choice of the native tones above chance (20%) into three ranges: Low spanned 20%-46%, Medium 47%-74%, and High 75%-100%. For the category-goodness ratings, the scale was divided into three ranges: Low = 1-2.9, Medium = 3-4.9, and High = 5-7.

Under both memory loads, T21 was Categorised as V22; T241 was Categorised as V44; T315 was Categorised as V214 (see Table 7.3). For all three tones, percent choices were in the high range and goodness ratings varied from medium to high range[26], suggesting moderate residual phonetic sensitivity. T33 was $\text{UnCategorised}_{\text{clustered}}$ and assimilated to V44 and V22 under low memory load, suggesting weak native language influences and was Categorised as V22 under high memory load, suggesting strong native language influences. In both cases, percent choices were in the medium range whereas ratings were in the high range, suggesting moderate residual phonetic sensitivity, which should moderately facilitate imitation. T45 was also $\text{UnCategorised}_{\text{clustered}}$ and assimilated to V35, V21, V214 under low memory load with percent choices for these response categories among low to medium range and ratings in the medium range, suggesting weak phonological constraints and moderate residual phonetic sensitivity to difference between native and non-native tone. Imitation in this case should be phonetically accurate and less susceptible to native constraints. But under high memory load, T45 was Categorised as V21 with percent choice and ratings in the moderate range, suggesting moderate to high native phonological influences and resulting in imitation affected by native phonological system.

[26] Memory load was manipulated only in the categorisation task but not the rating task that follows. Thus, since memory load is a between-subject factor, rating differences could reflect differences between two groups.

Table 7.3 Assimilation of Thai tones into Vietnamese tone categories under low versus high memory loads (from Chapter 6). Categories in bold are choices that were significantly above chance: 20% for Vietnamese; "*" = Categorised tone. Assimilations: C = Categorised, U = UnCategorised. Ratings: 1 = Poor, 7 = perfect; mean ratings are displayed. "-" = no response.

	Thai	T45		T33		T21		T315		T241	
	Response	%	rating	%	rating	%	rating	%	rating	%	rating
Low memory	SV44	6.3	3.5	**43.5**	5.4	4.3	3.6	0.2	1	**82.7***	5.7
	SV22	2.2	2.4	**51.8**	5.3	**88.7***	5.4	1.8	3.4	13.2	5.4
	SV35	**24.6**	4.7	1.1	2.2	0.5	4.5	7.6	5.8	1.8	3.5
	SV21	**42.7**	4.3	3.6	3.6	6.3	3.9	5.6	3.5	1.6	3.8
	SV214	**24.1**	4.3	-	-	0.2	1	**84.8***	5.2	0.7	3.3
Assimilation		$U_{clustered}$		$U_{clustered}$		C		C		C	
High memory	SV44	0.4	1	**38.1**	5	3.4	3.1	0.2	1	**81.1***	4.8
	SV22	0.4	2.5	**61.6***	5	**93.7***	4.7	-	-	16.6	4.5
	SV35	15.4	3.9	-	-	-	-	1.6	4.3	1.8	3
	SV21	**60.9***	4	0.2	1	1.6	2.3	6.8	2.8	0.4	1.5
	SV214	22.8	3	-	-	1.3	2.5	**91.4***	4	-	-
Assimilation		C		C		C		C		C	

7.7.2 Results

As in Experiment 1, the signed deviation scores for duration, $F0_{excursion}$, $F0_{mean}$, and $F0_{maxloc}$ were each selected as a dependent variable and fitted with a separate linear mixed-effects model. Memory load (low vs. high), talker variability (constant vs. variable), vowel variability (constant vs. variable) and Thai tone (T1-T5) were used as fixed factors. We first ran the analysis with participants as a random factor including random slopes for all within subject factors, i.e., talker

and vowel variability (as suggested by Barr et al., 2013). The models converged but were too complex to estimate *p* values for the fixed effects of interest. Thus, we dropped the random slopes and participants was specified as a random intercept in each model. Four models were built to test all main effects and interactions. To calculate the *p*-values for the fixed effects, we again used the Kenward-Roger degrees of freedom approximation (Halekoh and Hojsgaard, 2014), and the *Anova* function from the *car* package in R, with test specified as "F" (see Table 7.4, and for statistical details see Appendix C, Table C.9-C.12).

Main effects of vowel variability on duration, talker variability on $F0_{maxloc}$ deviation scores and tone type in all four acoustic-related deviation scores were found but the main effect of memory load was non-significant for all deviation scores (see Table 7.4). Imitation in term of duration deviations was more accurate in constant (*M* = .015, 95% CIs [.011, .020]) than variable vowel blocks (*M* = .022, 95% CIs [.017, .026]) as predicted. Imitation in term of $F0_{maxloc}$ deviations was unexpectedly less accurate in constant (*M* = -.084, 95% CIs [-.090, -.077]) than variable talker blocks (*M* = -.005, 95% CIs [-.015, .006]), which should be interpreted with respect to tone types given the significant interactions between talker variability and tone types.

To further examine the main effect of tone types, we ran multiple comparisons with Tukey adjustments to test the pairwise differences among tone types (see Appendix C, Table C.13 for statistical details). For deviation scores on duration, all pairwise comparisons were significant except for that between T241 (*M* = .053, 95% CIs [.047, .060]) and T33 (*M* = .061, 95% CIs [.054, .067]). The confidence interval of the deviation scores for T21 (*M* = -.003, 95% CIs [-.011, .005]) indicated the least deviation and the best imitation. T315 (*M* = -.036, 95% CIs [-.044, -.029]) and T45 (*M* = .018, 95% CIs [.011, .024]) were imitated with moderate deviations.

Table 7.4 Model details of acoustic measure deviation scores of Vietnamese imitators. Significant effects are in bold; marginal effects are in parentheses.

	Duration			$F0_{mean}$			$F0_{excursion}$			$F0_{maxloc}$		
	F	*df*	*p*	*F*	*df*	*p*	*F*	*df*	*p*	*F*	*df*	*p*
Memory (Mem)	0.0	1, 30	0.977	0.3	1, 30	0.580	0.7	1, 30	0.398	0.1	1, 30	0.781
Talker (Tlk)	2.1	1, 5019	0.146	0.2	1, 5019	0.644	0.5	1, 5019	0.469	205.0	1, 5019	**<.001**
Vowel (V)	3.7	1, 5019	**0.054**	1.7	1, 5019	0.187	0.1	1, 5019	0.793	0.2	1, 5019	0.631
Tone	130.0	4, 5019	**<.001**	296.9	4, 5019	**<.001**	578.0	4, 5019	**<.001**	208.4	4, 5019	**<.001**
Mem×Tlk	1.0	1, 5019	0.319	1.7	1, 5019	0.192	0.5	1, 5019	0.499	0.4	1, 5019	0.545
Mem×V	0.7	1, 5019	0.412	5.7	1, 5019	**0.017**	0.0	1, 5019	0.844	0.0	1, 5019	0.912
Tlk×V	0.2	1, 5019	0.679	0.4	1, 5019	0.533	1.5	1, 5019	0.215	0.5	1, 5019	0.468
Mem×Tone	7.0	4, 5019	**<.001**	8.0	4, 5019	**<.001**	6.4	4, 5019	**<.001**	1.9	4, 5019	0.115
Tlk×Tone	10.3	4, 5019	**<.001**	13.4	4, 5019	**<.001**	5.0	4, 5019	**0.001**	62.4	4, 5019	**<.001**
V×Tone	0.7	4, 5019	0.573	0.6	4, 5019	0.681	0.4	4, 5019	0.818	1.0	4, 5019	0.384
Mem×Tlk ×V	3.2	1, 5019	(0.072)	0.1	1, 5019	0.776	0.0	1, 5019	0.854	0.1	1, 5019	0.793
Mem×Tlk×Tone	0.6	4, 5019	0.637	1.7	4, 5019	0.140	1.1	4, 5019	0.335	0.0	4, 5019	0.999
Mem×V ×Tone	0.3	4, 5019	0.868	0.7	4, 5019	0.603	1.1	4, 5019	0.378	0.4	4, 5019	0.804
Tlk×V×Tone	0.2	4, 5019	0.931	0.5	4, 5019	0.717	0.3	4, 5019	0.860	1.9	4, 5019	0.104
Mem×Tlk×V×Tone	0.7	4, 5019	0.611	0.8	4, 5019	0.508	0.3	4, 5019	0.877	0.2	4, 5019	0.959

For deviation scores on $F0_{mean}$, all pairwise comparisons were significant except for the comparison between T241 (*M* = -.021, 95% CIs [-.024, -.017]) and T33 (*M* = -.024, 95% CIs [-.027, -.021]). T315 was imitated with the largest positive $F0_{mean}$ deviation score (*M* = .047, 95% CIs [.042, .051]), indicating higher $F0_{mean}$ than the target stimuli, and the imitation was the most deviant in this respect. On the other hand, T45 (*M* = -.038, 95% CIs [-.042, -.033]), T33, T241 were all imitated with negative $F0_{mean}$ deviation scores, indicating lower $F0_{mean}$ than the stimuli, and all with moderate deviations. The $F0_{mean}$ deviation scores for T21 (*M* = -.0004, 95% CIs [-.003, .003]) was the smallest, indicating the best imitation.

For deviation scores on $F0_{excursion}$, T315 (*M* = .244, 95% CIs [.231, .256]), T33 (*M* = .095, 95% CIs [.089, .101]), T45 (*M* = .056, 95% CIs [.049, .064]), and T241 (*M* = .027, 95% CIs [.019, .035]) were imitated with positive deviation scores, indicating larger F0 excursion than the stimuli. T315 was imitated with the largest excursion and thus was the least accurately imitated in this respect, followed by T33, T45 and T241. The $F0_{excursion}$ deviation scores for T21 (*M* = .001, 95% CIs [-.005, .007]) were the smallest, indicating the best imitation.

For $F0_{maxloc}$, all the pairwise comparisons were significant. T315 was the only tone imitated with a positive deviation score (*M* = .082, 95% CIs [.063, .101]), indicating later F0 maximum location than the stimuli. T241(*M* = -.143, 95% CIs [-.154, -.133]), T45 (*M* = -.096, 95% CIs [-.109, -.083]), T33 (*M* = -.069, 95% CIs [-.082, -.056]), were imitated with negative deviation scores, indicating earlier F0 maximum location than the stimuli. T241 was imitated with the largest deviation and the least accuracy, followed by T45 and T33. The $F0_{maxloc}$ deviation scores for T21 (*M* = .006, 95% CIs [-.002, .015]) indicated the smallest deviation and the best imitation.

In addition to main effects, to break down memory load × tone type interactions for syllable duration, $F0_{mean}$, and $F0_{excursion}$, we did pairwise comparisons with Tukey adjustments (see

Appendix C Table C.14 for statistical details). We report here only the most important significant differences for our predictions: comparisons between the same tone type under different memory loads. Only T315 showed significant differences between the two memory loads. Both duration and $F0_{mean}$ deviation scores for T315 indicated more accurate imitation in low ($M_{duration}$ = -.022, 95% CIs [-.032, -.013]; M_{F0mean} = .039, 95% CIs [.033, .045]) than high memory load ($M_{duration}$ = -.051, 95% CIs [-.061, -.040]; M_{F0mean} = .054, 95% CIs [.048, .060]), as predicted.

Similarly, to examine talker variability × tone type interactions for all four acoustic measures, we ran pairwise comparisons with Tukey adjustments (see Appendix C Table C.15 for statistical details). We report here only the significant differences for comparisons between the same tone type under different talker variability conditions.

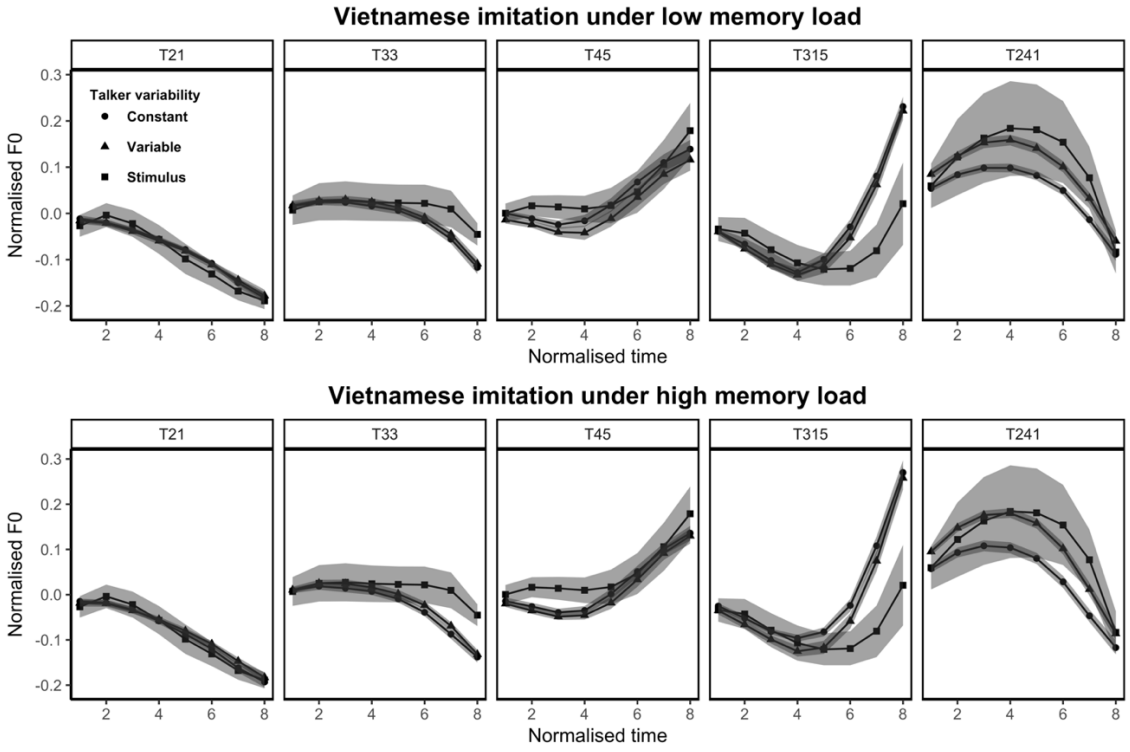


Figure 7.5 The time-and-Lobanov-normalised mean F0 contours of the Thai stimulus tones and their imitations produced by Vietnamese participants. Ribbons indicate 95% confidence intervals.

Imitation of $F0_{maxloc}$ for T315 was more accurate in constant (M = -.027, 95% CIs [-.039, .016]) than variable talker blocks (M = .190, 95% CIs [.157, .224]). However, that of $F0_{maxloc}$ for T33 and T241 was less accurate in constant (M_{T33} = -.133, 95% CIs [-.149, -.117]; M_{T241} = -.179, 95% CIs [-.192, -.165]) than variable talker blocks (M_{T33} = -.005, 95% CIs [-.023, .013]; M_{T241} = -.108, 95% CIs [-.123, -.093]). However, given that T33 was imitated as a level tone and the F0 range in T33 imitations was small, the difference in $F0_{maxloc}$ deviation scores is not very meaningful. Earlier realisation of $F0_{maxloc}$ in the imitation of T241 will render the contour more like a falling tone (see Figure 7.5). Mandarin imitators also showed this pattern T241.

Duration deviation scores for T21 were more accurate imitation in constant (M = .018, 95% CIs [.007, .029]) than variable talker blocks (M = -.024, 95% CIs [-.034, -.014]). Imitation in terms of $F0_{mean}$ for T241 were more accurate in constant (M = -.009, 95% CIs [-.014, -.004]) than variable talker blocks (M = -.033, 95% CIs [-.038, -.027]).

We ran multiple comparisons with Tukey adjustments to test the pairwise differences for the vowel variability by memory load interaction in $F0_{mean}$. Under high memory load, imitation in terms of $F0_{mean}$ were more accurate in constant (M = -.005, 95% CIs [-.009, -.001]) than variable vowel blocks (M = -.011, 95% CIs [-.015, -.008]), t = -2.621, df = 5019.1, p = .0436. All other comparisons were not significant.

7.7.3 Discussion

First, memory load showed no main effect but interactions with tone types for syllable duration, $F0_{mean}$, and $F0_{excursion}$. However, similar to Mandarin imitators, only imitation of T315 showed differences between two memory loads. Imitation of T315 in terms of syllable duration and $F0_{mean}$ was more accurate under low memory load, when rich phonetic details of F0 properties are still available and a phonetic mode of perception is activated, than under high memory load as we

expected. In addition, when rich phonetic details have decayed and imitators use a phonological mode of perception under high memory load and imitation in terms of duration was shorter than the stimulus reflecting native phonological constraints as the native category that T315 was assimilated to, i.e., V214, also has a shorter duration than T315.

Second, imitation was more accurate in constant than variable talker blocks but this effect was limited to some tones and in some deviation scores. Deviation scores of $F0_{maxloc}$ for T315, of syllable duration for T21 and of $F0_{mean}$ for T241 indicated more accurate imitation in constant than variable talker blocks. These results support our hypothesis that in constant talker blocks listeners used a phonetic mode of perception and were more sensitive to specific, concrete temporal and F0 properties of the stimuli and consequently imitated the stimuli more accurately than in variable talker blocks, where they appear to have used a phonological mode of perception. In a phonological mode of perception, native phonological perceptual routines were activated, constraining imitation. T315 was Categorised as V214, which has a larger $F0_{maxloc}$. Compatibly, imitation of $F0_{maxloc}$ for T315 was larger in variable than constant talker blocks. T21 was Categorised as V22 which has a shorter syllable duration. Syllable duration of T21 in imitation was shorter in variable than constant talker blocks. Similarly, T241 was Categorised as V44, which has a lower $F0_{mean}$, and imitation of T241 had a lower $F0_{mean}$ in variable than constant blocks. These findings support our hypothesis that imitation should reflect more native language influence in variable than constant talker blocks.

Third, there was a significant main effect of vowel variability on duration. Imitation was more accurate in terms of duration in constant blocks when listeners used a phonetic mode of perception than in variable talker blocks when low level phonetic information was variable, and listeners used a phonological mode of perception. In additional, the effect of vowel variability on $F0_{mean}$ was

significant only when the memory load is high. Under high memory load, imitation of $F0_{mean}$ was more accurate in constant than variable vowel blocks. We argue that under high memory load, listeners used a phonological mode of perception, reducing their ability to process concrete F0 properties of the stimuli. Consequently, imitation was less accurate in variable vowel blocks, where imitators have to abstract phonological pitch features from the more variable stimulus, than constant vowel blocks.

Fourth, imitation by Vietnamese participants reflected the unique native language influence as indicated by their perceptual assimilation patterns. Categorised assimilation indicates strong native language influence and should constrain native imitation. T315 was Categorised with a high percent choice into V214, which has a larger $F0_{excursion}$ and later $F0_{maxloc}$ than T315 (see Table 7.3). T315 was imitated with larger $F0_{excursion}$ and later $F0_{maxloc}$, like V214. Similarly, T241 was Categorised as V44 with high percent choice, which is lower than T241 in $F0_{mean}$. T241 was imitated with lower $F0_{mean}$ than the target tone, and more like the native tone to which it had been assimilated, as we predicted. In both cases, high percent choice in categorisation indicates low residual sensitivity to phonetic differences between non-native tones and native tones, and should lead to deviations that resemble the corresponding native tones.

On the other hand, for non-native tones that are Categorised with moderate percent choice and/or good ratings, listeners should display moderate residual phonetic sensitivity in perception and imitation. In the previous perception study (see Table 7.3), the same Vietnamese listeners as in the present study Categorised T21 into V22 with high percent choice, but the goodness rating was in the medium range under high memory load, suggesting moderate residual sensitivity to the difference between T21 and V22. V22 has shorter duration and smaller $F0_{excursion}$ than T21, but the imitation of T21 had the smallest deviation of all the Thai tones, showing little deviation from the

native Thai stimuli on all four measures. This suggests that phonological influence from the native Vietnamese tone was moderate and Vietnamese imitators detected the phonetic details of the target stimuli and realised them accurately in imitation.

7.8 General discussion

The two experiments reported here showed that cognitive factors shifted imitators' modes of perception and affected imitation. Native phonological and phonetic factors as indicated in the perceptual assimilation results in Chapter 6 predicted how non-native tone imitation was influenced by native languages. In this section, effects of memory load and talker/vowel variability on non-native imitation will be discussed first and followed by considerations of native language phonological and phonetic effects.

7.8.1 Effects of memory load

According to principles of PAM and ASP, under low memory load, listeners retain rich phonetic details of F0 properties in working memory and thus are biased toward using a phonetic mode of perception. In this mode, listeners attend more to phonetic details and are less constrained by native language phonological pitch features. On the other hand, under high memory load, phonetic details decay in working memory, and a phonological mode of perception should be activated. In the phonological mode, listeners are less sensitive to phonetic details because they have faded, thus they have only more lasting perceptual abstractions available in memory. Consequently, their native language phonological system has more influence. Extending these principles to imitation, it should be phonetically more accurate under low than high memory load.

While vowel (Repp & Williams, 1985) and segmental length (Asano & Braum, 2016) imitations have been found to be insensitive to memory load manipulations, our study on non-native tone

imitation by Mandarin and Vietnamese participants did find some evidence of memory load effect mostly for T315. Indeed, for Mandarin imitators, $F0_{mean}$ for T315 were imitated more accurately under low than high memory load. For Vietnamese imitators, similarly, imitation of syllable duration and $F0_{mean}$ for T315 was more accurate under low than high memory load. These results are consistent with our hypothesis that listeners should be biased toward a phonetic mode of perception under low memory load, when phonetic details in short-term memory are still available, and thus to imitate non-native F0 properties more accurately. Under high memory load, however, we reason that the rich array of fine-grained phonetic details in short-term memory will have faded, shifting imitators toward a phonological mode of perception. In this mode, native phonological perceptual routines should be more activated, biasing listeners to imitate Thai tones with abstract phonological features of native tones. In line with this, Vietnamese listeners Categorised T315 as V214, which is shorter than T315, and their imitations of T315 were shorter under high than low memory load, consistent with a phonological influence from native tone V214.

The fact that memory load effects were limited mostly to a phonetically complex falling-rising tone suggests that for simple tones, i.e., level, rising or falling, their phonetic details are less susceptible to decay in memory and thus are easy to imitate. Our manipulation of memory load may not be sensitive enough to test this idea. To test our hypothesis, future research can add a secondary task before imitation, such as counting digits, and this may substantially increase memory load.

7.8.2 Effects of talker and vowel variability

In general, our prediction that imitation is more accurate in constant than variable talker blocks is also supported by results from both language groups. For both Mandarin and Vietnamese imitators, $F0_{maxloc}$ for T315, syllable duration for T21, and $F0_{mean}$ for T241 were more accurately imitated in

constant than variable talker blocks. We argue that in constant talker blocks, where specific F0 properties are constant and reliable, imitators will use a phonetic mode of perception and focus more on phonetic level information. Consistent with this reasoning, their imitation was phonetically accurate. On the other hand, in variable talker blocks, we posited that imitators would be biased to use a phonological mode of perception because the low level phonetic information is variable, and they should therefore rely on abstract phonological temporal and pitch features via their native phonological perceptual routines. In non-native imitation, the phonological mode should result in imitation that is less phonetically accurate and more constrained by native phonological features. This hypothesis is supported by our findings that imitation in variable talker blocks displayed features from native tones.

Vowel variability appears to be easier to process than talker variability, as it affected only a small number of measures and only for Vietnamese participants. The Vietnamese group, under high memory load, imitated $F0_{mean}$ more accurately in constant than variable vowel blocks. Duration of their imitation was also more accurate in constant than variable vowel blocks. For Mandarin imitators, vowel variability did not yield any main effects or interactions for any of the deviation scores we examined.

In Chapter 6, we had found no significant effects of talker or vowel variability in categorisation for both Mandarin and Vietnamese groups, but discrimination was more accurate in constant talker/vowel blocks than variable talker/vowel blocks for both language groups. We reasoned that discrimination requires listeners to compare two non-native phones phonetically, which can be accomplished via phonetic level processing and thus listeners focused more on details at the phonetic level. In this sense, imitation is more similar to discrimination than to categorisation in that imitators need to reproduce the phonetic details of the non-native tone and consequently must

attend to the phonetic details of the target stimuli, which is hindered by talker/vowel variability because it biases toward the phonological mode of perception.

7.8.3 Effects of native language phonological and phonetic factors

Deviations in imitation can be partially accounted for by native language phonological constraints as indicated by assimilation types in perception, whereas phonetic imitation accuracy is commensurate with residual phonetic sensitivity in perception as indicated by percent choice and goodness ratings. For Categorised tones with high percent choice and/or goodness ratings, native phonological influences constrained non-native imitation, resulting in deviations similar to characteristics of the assimilated corresponding native tones. For Mandarin imitators, deviations in their imitations of T241 and T45 were compatible with the native Mandarin tones that they were assimilated to. Similarly, for Vietnamese imitators, T315 was affected by the native category that it was assimilated to, i.e., V214.

Even when a non-native tone was phonologically Categorised as a native tone, residual phonetic sensitivity to the differences between native and non-native tones can facilitate imitation of non-native tones. T21 was Categorised into M51 by Mandarin listeners and into V22 by Vietnamese listeners in Chapter 6 but listeners of both language groups showed moderate residual phonetic sensitivity to differences between native and non-native tones. T21 was the best imitated tone with low deviation scores in all measures for both language groups and not affected by the native tone it was assimilated to. This suggests that imitators retained phonetic details of the target stimuli and instantiate them in their imitations.

There was only one type of UnCategorised assimilation, i.e., $\text{UnCategorised}_{\text{clustered}}$ in the present study. Non-native tones of this type should bear weak native phonological effects as listeners did not perceive strong phonological similarity to any single native category but weak similarities to

two or three native categories. Consequently, their imitation should not be affected by any single native category. For Mandarin imitators, T315 was split between M35 and M214. Its deviations in imitation cannot be attributed to either M35 and M214. For Vietnamese imitators, assimilation of T45 and T33 were $\text{UnCategorised}_{\text{clustered}}$ under low memory load, reflecting weak phonological influence, but were Categorised under high memory load, indicating strong phonological influence. However, comparisons of the two tones across two memory loads were not significantly different for any deviation measures. Given that imitation requires phonetic level details, and percent choice and goodness ratings of native response categories under both memory load were in the low-to-medium range and thus comparable, we speculate that the comparable phonetic sensitivity outweigh differences in phonological effects as indicated by difference assimilation types.

Mandarin and Vietnamese imitators have different native tone systems with different phonetic realisations of each native phonological categories. These differences clearly affected some aspects of their imitations as reflected in the comparison of the same tone imitated by the two language groups (see Table C.16 in Appendix C for full statistical details). For example, T21 was Categorised as M51 by Mandarin participants, which has a larger $\text{F0}_{\text{excursion}}$ than V22, which the Vietnamese participants assimilated T21 to. Although T21 was among best imitated tones for both Mandarin and Vietnamese group, $\text{F0}_{\text{excursion}}$ was larger in the Mandarin (M = .053, 95% CIs [.045, .061]) than Vietnamese imitation (M = .001, 95% CIs [-.005, .007]). Similarly, Mandarin participants Categorised T241 to M51, which has a larger $\text{F0}_{\text{excursion}}$ than the Vietnamese high level tone V44 to which Vietnamese groups had Categorised T214. Although T241 was imitated with larger $\text{F0}_{\text{excursion}}$ than the stimuli by both language groups, it was larger for Mandarin (M = .102, 95% CIs [.089, .114]) than Vietnamese participants (M = .027, 95% CIs [-.035, .019]). In these cases, imitations of Categorised non-native tones were affected by each group's native tone

features, supporting the extension of the PAM principle that Categorised assimilation has strong native phonological influence on imitation performance.

7.9 Conclusion

In conclusion, cognitive factors, i.e., memory load, talker and vowel variability, affected mode of perception and consequently affected accuracy in non-native imitation of lexical tones. Imitations were less accurate and more constrained by native language phonological features under high memory load and in variable talker/vowel blocks where imitators used a phonological mode of perception and had low sensitivity to concrete temporal and F0 properties. On the other hand, under low memory load and in constant talker/vowel blocks, imitators used a phonetic mode of perception in which they showed more sensitivity to concrete temporal and F0 properties and produced phonetically more accurate imitations. Deviations in imitation among different tones can be attributed to native phonological influences as indicated by their perceptual assimilation patterns, in line with PAM-based predictions and supporting the articulatory commonality between non-native perception and imitation assumed by PAM. Although non-native listeners Categorised non-native tones into their native categories, they could also retain moderate residual sensitivity to phonetic details indicated by percent choice and goodness ratings. When imitating non-native tones, they used this residual sensitivity to produce phonetically more accurate imitations. The current findings thus have substantive implications for theories of perception and production. Phonetic accuracy in non-native imitation is commensurate with the amount of phonetic sensitivity in perception. Deviation in non-native imitation can be at least partially traced to native phonological constraints, which are predictable from perceptual assimilation patterns. Native language phonological constraints and residual phonetic sensitivities as well as phonological and phonetic modes of perception should be considered when researching non-native tone imitation

and learning. On the applied side, the results suggest that teachers of tone languages should tailor their pedagogy to address potential problems caused by native phonological influences for students of different language backgrounds.

Chapter 8. General discussion

This thesis has explored how native phonological and phonetic factors interplay with cognitive factors in non-native lexical tone perception and imitation. It has been found that native tone categories affected non-native perceptual assimilation, discrimination and imitation of Thai tones by native speakers of two tonal languages, Mandarin and Vietnamese. Both perception and imitation processes were modulated by cognitive factors, such as memory load, talker and vowel variability as they biased listeners/imitators to a phonological versus phonetic mode of perception.

This thesis makes three novel contributions to the field, filling knowledge gaps identified in Chapter 4. First, it examined non-native tone perception, which is an understudied phenomenon relative to consonants and vowels, by native tone language listeners/speakers, whose native language influences can operate at both phonological and phonetic level, c.f., non-tonal speakers, specifically Mandarin and Vietnamese, the latter group being understudied.

Second, non-native tone perception, both assimilation and discrimination, and imitation are considered dynamic processes as cognitive factors render listeners/imitator to shift between phonological and phonetic mode of perception/imitation. The interplay between cognitive and native phonological/phonetic factors offers a more comprehensive picture of speech perception and imitation.

Third, the imitation experiment linked perception and production. By including perception and imitation in a single project, it is possible to compare imitation performance with perceptual assimilation patterns to examine native language influences.

The thesis contains three experimental chapters, with Chapters 5 and 6 examining speech perception and Chapter 7 imitation. In Chapter 5, three groups of tone language listeners, i.e., Mandarin and Vietnamese including Northern and Southern dialects, were tested in a categorisation task. Chapter 6 examined discrimination of five non-native tone contrasts selected based on assimilation patterns reported in Chapter 5 and categorisation of five Thai tones by Mandarin and Southern Vietnamese listeners under different systematically manipulated cognitive conditions. The same participants imitated five Thai tones under analogous cognitive factor manipulations in Chapter 7. The main findings of each of these three experimental chapters are discussed below followed by consideration of their contribution to theories of non-native speech perception and production. It should be noted that individual variation did exist in the assimilation, discrimination, and imitation experiments. Previous studies have shown that inter-individual differences in the perceptual assimilation of vowels are particularly high (Tyler et al., 2014). The same applied here with tone assimilation. A detailed analysis of individual differences requires a larger number of participants for each group and a larger number of trials for each participant than in the present studies, and falls beyond the scope of the thesis, but would be interesting and important to address in further work on non-native tone assimilation. Nevertheless, for all perception and imitation results, I reported 95% confidence interval so that readers can get a sense of the extent of individual variation.

8.1 Summary of findings

8.1.1 Non-native tone perception

The experiments in Chapters 5 and 6 have examined perceptual assimilation of five Thai tones by Mandarin and Vietnamese listeners who had no experience with Thai. First and foremost, the results showed that both native phonological and phonetic factors affected non-native tone

assimilation by tone language listeners. The fact that most Thai tones were assimilated into native tone categories indicates strong native phonological categorisation effects. On the other hand, phonetic differences between non-native and native tones are reflected in different ranges of percent choice and goodness rating. For example, T33 was Categorised as M55 by Mandarin listeners, which reflected strong native phonological influences. Listeners also showed residual phonetic sensitivity to differences between native and non-native categories by a relatively low Category-goodness rating. What is more interesting in this case is that the percent choice of T33 assimilation as M55 was in the high range (>75%) but the rating was in the medium range (< 5). We argue that this discrepancy reflects the differences between percent choice and goodness rating in quantifying residual phonetic sensitivity. Percent choice reflects consistency of responses within a categorisation task and thus is a sign of phonetic sensitivity during the categorisation process where the focus is on fitting non-native phones into abstract native phonological categories; on the other hand, goodness ratings are the product of a rating task where the focus is on evaluating phonetic variations *within* a phonological category.

In addition, phonological and phonetic differences between two regional varieties of a language have impacts on non-native assimilation. Phonologically, Southern Vietnamese has a tone merger (i.e., SV214) of two tones, namely, NV214, NV415 that are distinct in Northern Vietnamese. Phonetically, NV415 has an extremely low dip in the middle, which is absent in Thai and Southern Vietnamese. SV21 but not NV21 has a final rising. These phonological and phonetic differences led to different assimilation patterns of two Thai phonetically rising tones T45 and T315 between Northern and Southern Vietnamese listeners. T45 was Categorised as NV35 but $\text{UnCategorised}_{\text{clustered}}$ and split among SV35, SV214 and SV21. T315 was Categorised as NV35 but as SV214. This is analogous to previous findings that phonological and phonetic

similarity/dissimilarity can affect the perception of approximants (Best & Strange, 1992; Bohn & Best, 2012; Hallé et al., 1999).

Another interesting finding of native language effects on categorisation is that strong phonological influence in Categorised assimilations resulted in shorter categorisation response time compared with UnCategorised assimilation which reflects weak native phonological influence. We reason that when there are no strong similarities to any native tones, UnCategorised assimilation incurs an extra processing cost caused by perceptual uncertainty due to phonological competition among and/or phonetic discrepancies from native categories.

As for discrimination, first of all, perceived similarity in the form of assimilation types successfully predicted the discrimination of non-native tone contrasts as outlined by PAM. When two non-native tones were Categorised to a single native tone category either equally good as in Single-Category or with different goodness ratings, as in Category-Goodness, the discrimination was less accurate than when two non-native tones were neatly Categorised into two different native tone categories as in Two-Category Non-overlapped or Two-Category/Non-overlap. For example, T33-T45 was a Two-Category assimilation by Mandarin listeners with no overlap in terms of native response categories and this contrast was better discriminated than T241-T21 which was a Category-Goodness assimilation.

Furthermore, perceived *phonological* (i.e., complete, partial and none) overlap in listeners' assimilations affected discrimination performance. Contrasts with complete phonological overlap in categorisation were discriminated more poorly than those with partial phonological overlap. For instance, T33-T21 (Two-Category/Non-overlap) was better discriminated by Mandarin listeners than T33-T241, which partially overlapped in the high memory load condition.

Apart from phonological overlap types, differences of residual phonetic sensitivity to deviation of native versus non-native tone categories also have a role to play in predicting discrimination performance particularly among contrasts of the same assimilation and phonological overlap types. To quantify that effect, two difference scores were calculated: the overlap score (Flege & MacKay, 2004; Levy, 2009) and the fit-index difference score (Wu et al., 2014). Both scores reflect phonetic sensitivity to each non-native tone in the contrast from the corresponding native tone. Both scores successfully predicted the better discrimination of T33-T45 than that of T33-T21, both of which were Two-Category assimilations with no phonological overlap (Two-Category/Non-overlap) for Mandarin listeners. This suggests that non-native contrasts that fall into the same phonological assimilation type could be discriminated differently and this difference can be predicted by the phonetic sensitivity to each non-native tone in the contrast from the corresponding native tone,

Lastly, memory load, talker and vowel variability, have affected non-native tone categorisation and discrimination differently in the series of perceptual experiments in Chapter 6. Assimilation responses showed more phonologically-based categorisation patterns, i.e., more Categorised assimilation and higher percent choice and lower goodness ratings under high memory load than low memory load but were unaffected by talker and vowel variability. In contrast, discrimination accuracy was reduced by talker and vowel variability but was unaffected by memory load for both Mandarin and Southern Vietnamese listeners.

8.1.2 Non-native tone imitation

The set of experiments reported in Chapter 7 have revealed that non-native imitation of the five Thai tones by Mandarin and Southern Vietnamese imitators, who also participated in perceptual experiments reported in Chapter 6, were affected by native language influence and cognitive factors.

First, memory load interacted with tone types for $F0_{mean}$, $F0_{excursion}$ and $F0_{maxloc}$ for Mandarin imitators, but for syllable duration, $F0_{mean}$, $F0_{excursion}$ for Vietnamese imitators. As we expected, $F0_{mean}$ deviations of T315 were smaller under low than high memory load for Mandarin imitators. For Vietnamese imitators, imitation of T315 was more accurate under low than high memory load in terms of syllable duration and $F0_{mean}$.

Talker variability had significant main effects on $F0_{maxloc}$ deviation scores and significant interactions with tone types for duration, $F0_{mean}$, and $F0_{maxloc}$ scores by Mandarin imitators. For Vietnamese imitators, there was a significant main effect of talker variability on $F0_{maxloc}$ scores as we found in the Mandarin group and significant interactions with tone types for all four acoustic measures. For both language groups, the $F0_{maxloc}$ in imitation of T315 was significantly more accurate in constant than variable talker blocks. Syllable duration for T21 and T45, and $F0_{mean}$ for T241, were also imitated more accurately in constant than variable talker blocks for Mandarin imitators. For Vietnamese imitators, syllable duration for T21, $F0_{mean}$ for T241 were imitated more accurately in constant than in variable talker blocks.

For vowel variability, a significant interaction between vowel variability and memory load for $F0_{mean}$ were found only for Vietnamese imitators. Under high memory load, imitation of $F0_{mean}$ was also more accurate in constant than variable vowel blocks. Both results are in line with the prediction that imitation should be more accurate in constant than in variable vowel blocks.

For both language groups, significant main effects of tone types were found for all four acoustic measures. Deviations of some tones could be accounted for by native language phonological constraints and residual phonetic sensitivities. For Mandarin imitators, T241 imitations had larger excursion and earlier F0 maximum location in imitations, consistent with the characteristics of M51 they had assimilated T241 to, suggesting clear L1 phonological influences. T45 was

assimilated as M35 (in Chapter 6), which is lower in $F0_{mean}$ than T45. The imitation of T45 also has lower $F0_{mean.}$ Similarly, for Vietnamese imitators, T315 was Categorised as V214 with high percent choice, which has larger $F0_{excursion}$ and later $F0_{maxloc}$. The imitation of T315 also has larger $F0_{excursion}$ and later $F0_{maxloc,}$ suggesting native language influence. The deviations in the imitation could be accounted for by phonological influence from the corresponding native tone.

For some other Thai tones which were also Categorised as native tones for both groups, the imitation was rather accurate. These Thai tones often have medium range percent choice and/or goodness ratings in categorisation, indicating that residual phonetic sensitivity to the target stimuli from native phonological categories. For Mandarin listeners, T21 was Categorised into M51 but percent choice and goodness ratings were in the medium range. Thus, the relative accurate imitation could be attributed to the moderate level of residual phonetic sensitivity to variations of the non-native tone from the native tone they assimilated to. Similarly, although T33 was perceptually assimilated into M55, which is higher in F0, $F0_{mean}$ of imitation of T33 was lower than the target. This indicates that participants still retained the phonetic details of F0 height when imitating T33. The same is true for Vietnamese imitators. T21 was Categorised into V22 with high percent choice, but the goodness rating was in the medium range under high memory load, suggesting moderate residual sensitivity to the difference between T21 and V22. The imitation of T21 had the smallest deviation of all the Thai tones, indicating weak phonological influence. When imitators retain moderate phonetic sensitivity to differences between native and non-native tones, they were able to detect phonetic details of the stimuli and instantiated them in imitation.

Despite of all these results, the effects of memory load and stimulus variability on non-native imitation were not equally strong for all tones but varies across different tones and different acoustic deviation scores measured. And I am not arguing that all these deviation score differences

will be perceptually meaningful, nor teachers should directly transform the results into teaching instructions. A perceptual evaluation experiment using native Thai listeners can further test the cognitive effects we found and a subsequent correlation analysis could reveal which acoustic deviations are more perceptually relevant.

8.2 Relevance of findings to non-native perception and production theories

The results of the experimental chapters presented in Chapters 5-7 shed new light on current non-native speech perception and production theories.

8.2.1 Native language influences on non-native tone perception

First and foremost, the basic PAM principle that non-native phones can be assimilated into native phonemes and their perceptual assimilation can predict the discrimination performance were upheld. When two non-native tones were Categorised to the same single native tone category (Single-Category or Category-Goodness), discrimination was less accurate than when two non-native tones were neatly Categorised into two different native tone categories (Two-Category). This further confirms that PAM can be extended to non-native tone perception especially for tone language listeners. Unlike many previous studies that compared tonal vs. non-tonal listeners, this thesis compares discrimination across two tone language groups and thus provides strong evidence for native tone system influences, especially for the comparisons of the same non-native tone contrast that was assimilated differently by each group. The observed discrimination accuracy confirms PAM predictions. T21-T33 is a good case in point. The T21-T33 contrast forms a Two-Category assimilation for Mandarin listeners but UnCategorised-Categorised/Single-Category assimilation for Southern Vietnamese listeners (as in Chapter 6). The PAM prediction that T21-T33 should therefore be better discriminated by Mandarin than Vietnamese listeners was upheld.

The findings have important implications for second language teaching of tone languages, as comparisons between two language groups clearly indicate that discrimination difficulties can vary as a function of students' first language backgrounds. Thus, teachers should design their pedagogy accordingly.

One novel theoretical contribution of this thesis regarding native language influence is the explication of native phonological and phonetic factors in non-native perception, both perceptual assimilation and discrimination, in line with PAM principles. In perceptual assimilation, Categorised versus UnCategorised assimilation types indicate native phonological effects whereas percent choice and category-goodness ratings showcase native phonetic effects. In addition, percent choice reflects residual phonetic sensitivity in the categorisation process where the attention focus is on assigning a native phoneme to the non-native phone. In contrast, goodness ratings mirror residual phonetic sensitivity in the rating process where the attention focus is on evaluating phonetic variations of the non-native phone from the native phoneme it is assimilated to. Three ranges (High, Medium and Low) were proposed to quantify percent choice and goodness ratings (see Chapter 5). It was expected and observed that in some cases, percent choice and ratings would not fall into the same range. For example, Mandarin listeners reliably Categorised the Thai mid level tone to their only native level tone, which has a different F0 height, indicating a strong phonological effect and with percent choice in the High range, indicating weak phonetic sensitivity in the categorisation process, but assigned it Low category-goodness ratings, indicating strong phonetic sensitivity in the rating process. These distinctions enrich and strengthen PAM's ability to make more accurate and detailed predictions that account for both native phonological and phonetic effects in perceptual assimilation.

As for discriminating non-native contrasts, the PAM framework (Best, 1995) was originally based

on assimilation types, i.e., Two-Category, Sing-Category, Category-Goodness and UnCategorised-Categorised and later considered (Faris et al., 2016, 2018) phonological overlap, i.e., Non-Overlap, Partially-Overlap, Completely-Overlap which count only above chance level native responses. The thesis added that phonetic overlap, in terms of overlap scores or fit index difference scores which include below chance level native responses (see Chapter 6), can help to predict variations of contrasts with the *same* assimilation and phonological overlap, which extends residual phonetic sensitivity to difference between native and non-native categories in assimilation to the case of discrimination. By extrapolating PAM principles regarding native phonological and phonetic effects for both perceptual assimilation and discrimination, the thesis lays the theoretical groundwork for further research that explores native influence on non-native perception.

8.2.2 Cognitive factors on non-native tone perception

Non-native speech perception is a dynamic process and cognitive factors demonstrably impact non-native tone perception, though different patterns were observed for categorisation and discrimination. The results support hypothesis based on ASP's principles that selective perception routines are activated to perceive native and non-native speech by detecting task-relevant information and listeners can switch between phonological and phonetic mode of perception as motivated by cognitive factors.

Memory load, by means of manipulating the temporal course of the judgements in the tasks, exerted different effects on categorisation and discrimination. Categorisation under high memory load was more phonological, reflected in more Categorised assimilation, and less sensitive to low level temporal and F0 differences between the non-native and native tone, reflected in higher percent choice as compared with that under low memory load. This is a novel finding, as no previous research has manipulated memory load in perceptual assimilation. High memory load

generally bias listeners to a phonological mode as compared with low memory load. This also suggests that care should be taken when comparing assimilation results across different studies as listeners may apply different modes of perception.

Discrimination was largely unaffected by memory load, compatible with previous findings on non-native perception of Mandarin and Cantonese tones (Lee et al., 1996) but different from previous findings with consonants (Asano, 2017; Werker & Tees, 1984; Werker & Logan, 1985). The cue-duration hypothesis (Fujisaki & Kawashima, 1970), which claims that phonetic details of longer-duration speech segments such as vowels are better retained in memory, can be extended to account for our findings. Tones, like vowels, extend over the entire sonorant part of the syllable, and thus are more enduring. In the AX task, the first stimulus should be stored in short-term memory longer and remain available for comparison with the second stimulus. This hypothesis can be further tested by using longer ISIs, e.g., 5 s or more. A previous research on bilingual perception of stop contrasts (Antoniou et al., 2012) also found difference between categorisation and discrimination, that is, manipulating the language mode shifted bilinguals' categorisation and goodness ratings, but not their discrimination.

The different effects observed for memory load on categorisation and discrimination could be accounted for by considering the different processes underlying these tasks. In the categorisation task, after hearing the stimulus, listeners could start the categorising process immediately but not in the discrimination task, where listeners have to wait for the second stimuli to start the comparing process, even though they were required to delay their response by a short or long interval. High memory load in the categorisation task lengthened the processing time so that listeners were more likely to engage in higher level phonological processing. Under low memory load, on the other hand, listeners were pushed to base their categorisation choice more on low level temporal and F0

property similarities. This hypothesis can be tested with a digit preload secondary task (Logan, 1979) in categorisation. This secondary task should interrupt processing before the response and add memory load at the same time.

Second, processing variability in speech has always been a central issue in speech perception research. The present study systematically manipulated talker and vowel context variability, which turned out to have different impacts on non-native tone categorisation and discrimination. Talker and vowel variability affected discrimination performance, whereas categorisation was largely intact. In constant talker/vowel blocks, discrimination was more accurate as listeners used a phonetic mode of perception and focused on phonetic details of temporal and F0 properties. On the other hand, listeners switch to a phonological mode in variable talkers/vowels blocks. Consequently, their discrimination accuracy was reduced as they attended to more perceived abstract pitch contours and heights and were more constrained by native language perceptual routines.

The differing effects on discrimination and categorisation may be attributed to the different levels of processing subserving the two tasks, as has been argued in previous studies (Antoniou et al., 2012, 2013). Categorisation tasks require phonological-level judgements and in this case these judgements are based on native phonological categories. Native tone language listeners should possess mechanisms to maintain phonological constancy in their native language. Thus, we speculate that when naïve listeners used their native language phonological categories, these mechanisms are also activated to process the non-native stimulus. In other words, perceivers assimilate not only non-native tones to native categories but also the phonetic properties of unfamiliar talkers to the key indexical features of their native speech community (Best, 2015). Conversely, in discrimination, listeners simply decide whether the tones in the stimulus pairs were

the same or not at lower and more phonetic level. Thus, listeners may base their choice on low level phonetic information about temporal and F0 properties and are more likely to be affected by stimulus variations.

In conclusion, categorisation and discrimination tasks are served by different underlying cognitive processes and are susceptible to different cognitive manipulations. High memory load but not high talker and vowel variability biased listeners to a phonological mode in categorisation whereas high talker and vowel variability but not high memory load biased listeners to a phonological mode in discrimination. Non-native speech perception theories should consider these effects together with native language phonological and phonetic effects when modelling variations in perceptual performance.

8.2.3 Native language influences on non-native tone imitation

One major theoretical question to address is whether or to what extent imitation deviations can be traced to native language influences and if these influences can be predicted by perceptual assimilation. SLM claimed that at least some errors in speech production have a perceptual basis but the theory did not explicate phonological or phonetic factors in the process. Given that PAM's assumptions about perception of articulatory information in speech, extending PAM principles to imitation, for non-native tones that are Categorised as native tones with high percent choice and goodness ratings, native phonological influences should be strong, and lead to inaccurate imitation. On the other hand, native language constraints should be weak for UnCategorised assimilations, where non-native phones cannot be assimilated into a single native category. The only subtype of UnCategorised assimilation found in the present study is UnCategorised$_{\text{clustered.}}$ In those cases, phonological influences were weak and deviations of imitation cannot be attributed to any single native category, unlike those in Categorised cases.

Within those phonological constraints, if percent choice and goodness ratings are small, residual sensitivity to within-category phonetic variations of the non-native phones from their native categories should be strong. Strong phonetic sensitivity should facilitate more accurate imitation of non-native tones.

Mandarin and Vietnamese imitators have different native tone systems and consequently different assimilation patterns of Thai tones. Comparing imitation deviations of the same tone across two language groups confirmed influence of the native language. For example, T21 was Categorised by Mandarin participants as M51, which is larger in $F0_{excursion}$ than V22 which Vietnamese participants assimilated T21 to. $F0_{excursion}$ was significantly larger in Mandarin imitation of T21 than Vietnamese imitation of T21. Similarly, T241 was Categorised as M51 which is larger in $F0_{excursion}$ than V44 which Vietnamese participants assimilated T241 to. T241 was also imitated with larger $F0_{excursion}$ by Mandarin than by Vietnamese participants. These results suggested that for Categorised non-native tones, native language constrained non-native imitation.

However, even for Categorised tones, listeners can still retain phonetic sensitivity in imitation as indicated by percent choice/goodness ratings. For example, T33 was phonologically Categorised as M55, which is phonetically higher than T33 in F0 mean. However, the category-goodness rating for this assimilation was in the medium range (as shown in Chapter 5), suggesting residual phonetic sensitivity to the difference between native and non-native tones. Imitator noticed the phonetic differences between their native tone category and the target stimulus. Consequently, $F0_{mean}$ was imitated lower than the stimuli. I hypothesise that when participants Categorised a non-native tone to a native one, they still retain phonetic sensitivity to the differences between native and non-native tones. If a phonetic mode of perception is activated, they use phonetic details they perceive in the stimulus to reproduce it accurately and differently from their native target. Alternatively, if

a phonological mode is activated, they turn to perceptual routines attuned by their native language and their imitation is constrained by native language.

In summary, PAM principles regarding native language phonological and phonetic factors, as indicated by assimilation type as well as percent choice and/or goodness ratings respectively, can be extended to account for non-native imitation performance.

8.2.4 Cognitive factors on non-native tone imitation

Non-native tone imitation deviations were not only affected by native language influence at phonological and phonetic level but also modulated by cognitive factors. Cognitive factors biased imitators toward phonological versus phonetic mode of imitation. First, although memory load did not affect vowels (Repp & Williams, 1985) and segment length (Asano & Braum, 2016) imitation, it did affect non-native imitation of some tones by Mandarin and Vietnamese participants in Chapter 7. For both language groups, imitation under low memory load was more accurate than high memory load for some tones with complex F0 contours like T315. These results supported our hypothesis that under low memory load listeners were able to use a phonetic mode of perception when phonetic details of temporal and F0 properties in short-term memory were still available and thus imitate non-native stimuli more accurately. Under high memory load, however, the rich array of fine-grained phonetic details faded and imitators changed into a phonological mode, and consequently the imitation of these tones was more deviant from the target stimuli and more constrained by perceptual abstractions of pitch features in native language phonology. I speculate, however, that this memory load effect is most evident with complex contour tones such as T315 (falling-rising). On the other hand, I argue that for simple tone contours, i.e., level, rising or falling, the simpler phonetic details are less susceptible to memory decay, making them easier to imitate more accurately even under high memory load. Thus for them, high memory load does

not significantly reduce imitation accuracy.

Second, Mandarin and Vietnamese imitations were more accurate in constant than variable talker blocks. But vowel variability affected imitation by Vietnamese participants only when they imitated under high memory load, in which imitation was also more accurate in constant than variable vowel blocks. The results indicate that high talker and vowel variability biased imitators to a phonological mode because phonetic temporal and F0 properties are too variable in the speech signal. In a phonological mode, imitators attend more to abstract phonological pitch features and less to specific and variable phonetic details, resulting in lower accuracy in non-native imitation. Moreover, as vowel variability affected only a small number of measures and affected only Vietnamese participants, it is speculated that vowel variability is easier to process than talker variability.

Perceptual experiments in Chapter 6 did not find any significant effects of talker or vowel variability in categorisation but did find that discrimination was more accurate in constant talker/vowel blocks than variable talker/vowel blocks. Discrimination requires listeners to compare F0 properties of two phones at a phonetic level. When talker and vowel variability increases, listeners have to change to a phonological mode and rely on native phonological perception routines. These native perception routines were formed by native language experience and were not accurate in perceiving non-native tones. Thus, their discrimination accuracy was reduced. Given that imitation was affected by similar cognitive factors, i.e., talker and vowel variability, imitation is more similar to discrimination than to categorisation. Imitators need to reproduce temporal and F0 properties of the non-native tones and consequently had to attend to the phonetic details of the target stimuli, including talker/vowel variability.

Results of three experimental chapters generally support PAM and ASP principles and predictions.

Native language phonological and phonetic factors affect non-native assimilation, the patterns of which in turn can be used to predict discrimination and imitation. Cognitive factors bias listeners to phonological and phonetic mode and consequently influence discrimination and imitation. In the next section, I will propose some future directions in research similar topics.

8.3 Future directions

This thesis examined non-native tone perception and imitation by native tone language listeners with no experience with the unfamiliar language, Thai. To fully understand second language development in terms of lexical tones, future research should investigate perception and imitation/production of different Thai tones by Mandarin and/or Vietnamese listeners of differing proficiency. Both SLM and PAM-L2 have predictions for the development of non-native phonetic categories.

According to PAM-L2 (Best & Tyler, 2007), when an L2 phone is Categorised as a given L1 phonological category, no further perceptual learning is likely to happen. This can be beneficial to L2 learning when the native category is phonetically similar to the non-native category. However, Categorised assimilation could be counterproductive for L2 tone production accuracy if the non-native category is phonetically different from the corresponding native category.

For non-native tones that are not Categorised as any single L1 phonological category but are heard as being similar to several L1 categories, i.e., UnCategorised assimilation, PAM-L2 predicts that one or two new L2 phonological categories may be formed. Similarly, SLM claims that new categories could be formed for this type of L2 phone, and that if the new phonetic category matches that of native speakers of the L2, then the L2 sound will be produced accurately. On the other hand, the newly established category could also deviate from both native and the target language as learners may shift the new category from their native category.

PAM-L2 predictions differ from SLM in that PAM considers the comparative relationships within the interlanguage phonological system in addition to the similarity of a given L2 phone to the closest individual native phonetic category. If the UnCategorised L2 phones are assimilated into different sets of L1 phonemes with little overlap in the native categories chosen, PAM-L2 predicts that two or more new categories could be formed. But if the UnCategorised L2 phones are identified as similar to the same set of L1 phonemes, then only a single new category would be formed, and discrimination of the contrasting L2 tones may remain difficult (Best et al., 2019). Findings of Chapters 6 and 7 can offer potential Thai tone candidates (Categorised and/or UnCategorised) for testing with second language learners of Thai of different proficiency.

Second, discrete acoustic deviation scores were measured to quantify imitation performance in Chapter 7. Memory load effects and stimuli variability effects were limited to a few tone and some of their acoustic measures. Human perceptual judgements can be used to further test the strength of these effects and whether deviation in individual acoustic measures will lead to misidentification. In addition, with the development of machine learning algorithms, it is desirable to use classification algorithms, such as Linear Discriminant Analysis or Support Vector Machine and RandomForest models, which allow multiple acoustic correlates to be modelled across languages in a way analogous to native listener classifications.

In conclusion, this thesis makes novel and important contributions to the understanding of non-native tone perception and imitation by native tone language listeners/speakers. The results support PAM's predictions regarding native language influence on non-native tone perception and imitation. In addition, both perception and imitation were affected by cognitive factors as these factors bias listeners/imitation toward phonological and phonetic mode of perception/imitation. Categorisation was shifted by memory load but remain unaffected by manipulations of talker and

vowel variability. In contrast, discrimination was not influenced by memory load but were affected by talker and vowel variability. Imitation showed effects of both memory load and talker variability. A link between perception and production was supported.

References

Abramson, A. S. (1962). The vowels and tones of standard Thai: Acoustical measurements and experiments. *International Journal of American Linguistics*, *28*(2), 1–146.

Abramson, A. S. (1972). Tonal experiments with whispered Thai. In A. Valdman (Ed.), *Papers in linguistics and phonetics to the memory of Pierre Delattre* (pp. 29–55). The Hague: Mouton.

Abramson, A. S. (1975). The tones of Central Thai: Some perceptual experiments. In J. G. Harris & J. R. Chamberlain (Eds.), *Studies in Thai linguistics in honor of William J. Gedney* (pp. 1–16). Central Institute of English Language.

Abramson, A. S. (1976). Thai tones as a reference system. In T. W. Gething, J. G. Harris, & P. Kullavanijaya (Eds.), *Tai Linguistics in Honor of Fang-Kuei Li* (pp. 1–14). Chulalongkorn University Press.

Abramson, A. S. (1978). Static and dynamic acoustic cues in distinctive tones. *Language and Speech*, *21*(4), 319–325.

Alispahic, S., Mulak, K. E., & Escudero, P. (2017). Acoustic Properties Predict Perception of Unfamiliar Dutch Vowels by Adult Australian English and Peruvian Spanish Listeners. *Frontiers in Psychology*, *8:52*. https://doi.org/10.3389/fpsyg.2017.00052

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*. *ICPhS 2007*, 361–364.

Antoniou, M., Best, C. T., & Tyler, M. D. (2013). Focusing the lens of language experience: Perception of Ma'di stops by Greek and English bilinguals and monolinguals. *The*

Journal of the Acoustical Society of America, *133*(4), 2397–2411. https://doi.org/10.1121/1.4792358

Antoniou, M., Best, C. T., Tyler, M. D., & Kroos, C. (2010). Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. *Journal of Phonetics*, *38*(4), 640–653. https://doi.org/10.1016/j.wocn.2010.09.005

Antoniou, M., Best, C. T., Tyler, M. D., & Kroos, C. (2011). Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching. *Journal of Phonetics*, *39*(4), 558–570. https://doi.org/10.1016/j.wocn.2011.03.001

Antoniou, M., Tyler, M. D., & Best, C. T. (2012). Two ways to listen: Do L2-dominant bilinguals perceive stop voicing according to language mode? *Journal of Phonetics*, *40*(4), 582–594. https://doi.org/10.1016/j.wocn.2012.05.005

Antoniou, M., & Wong, P. C. M. (2015). Poor phonetic perceivers are affected by cognitive load when resolving talker variability. *The Journal of the Acoustical Society of America*, *138*(2), 571–574. https://doi.org/10.1121/1.4923362

Antoniou, M., Wong, P. C. M., & Wang, S. (2015). The Effect of Intensified Language Exposure on Accommodating Talker Variability. *Journal of Speech, Language, and Hearing Research*, *58*(3), 722–727. https://doi.org/10.1044/2015_JSLHR-S-14-0259

Asano, Y., & Braum, B. (2016). Does speech production in L2 require access to phonological representations? *Proceedings of the International Conference on Speech Prosody*, 237–241.

Asano, Y. (2017). Discriminating Non-Native Segmental Length Contrasts Under Increased Task Demands. *Language and Speech*, *61*(3), 409–429. https://doi.org/10.1177/0023830917731907

Baddeley, A. (2010). Working memory. *Current Biology*, *20*(4), R136–R140. https://doi.org/10.1016/j.cub.2009.12.014

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Bang, H.-Y., Sonderegger, M., Kang, Y., Clayards, M., & Yoon, T.-J. (2018). The emergence, progress, and impact of sound change in progress in Seoul Korean: Implications for mechanisms of tonogenesis. *Journal of Phonetics*, *66*, 120–144. https://doi.org/10.1016/j.wocn.2017.09.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Best, C. T. (1995). A direct realist view of cross-language speech perception. In Winifred Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.

Best, C. T. (2015). Devil or Angel in the Details?: Perceiving phonetic variation as information about phonological structure. In J. Romero & M. Riera (Eds.), *Phonetics-Phonology Interface: Representations and Methodologies* (pp. 3–31). John Benjamins Publishing Company

Best, C. T. (2019). The Diversity of Tone Languages and the Roles of Pitch Variation in Non-tone Languages: Considerations for Tone Perception Research. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00364

Best, C. T., Avesani, C., Tyler, M. D., & Vayra, M. (2019). PAM Revisits the Articulatory Organ Hypothesis: Italians' Perception of English Anterior and Nuu-Chah-Nulth Posterior Voiceless Fricatives. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. H. Sørensen (Eds.), *A Sound Approach to Language Matters: In Honor of Ocke-Schwen Bohn* (pp. 13–40).

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 345–360. https://doi.org/10.1037/0096-1523.14.3.345

Best, C. T., & Strange, W. (1992). Effects of Phonological and Phonetic Factors on Cross-Language Perception of Approximants. *Journal of Phonetics*, *20*(3), 305–330.

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, *109*(2), 775–794.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language Experience in Second Language Speech Learning* (pp. 13–34). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.17.07bes

Best, C. T., Tyler, M. D., Gooding, T. N., Orlando, C. B., & Quann, C. A. (2009). Development of Phonological Constancy: Toddlers' Perception of Native- and Jamaican-Accented Words. *Psychological Science*, *20*(5), 539–542. https://doi.org/10.1111/j.1467-9280.2009.02327.x

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–345.

Boersma, P. (1998). Functional phonology: Formalizing the interactions between articulatory and perceptual drives. [Doctoral Thesis, University of Amsterdam].

Bohn, O.-S., & Best, C. T. (2012). Native-language phonetic and phonological influences on perception of American English approximants by Danish and German listeners. *Journal of Phonetics*, *40*(1), 109–128. https://doi.org/10.1016/j.wocn.2011.08.002

Bohn, O.-S., & Flege, J. E. (1992). The Production of New and Similar Vowels by Adult German Learners of English. *Studies in Second Language Acquisition*, *14*(2), 131–158. https://doi.org/10.1017/S0272263100010792

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251. https://doi.org/10.1017/S0952675700001019

Brunelle, M. (2009). Tone perception in Northern and Southern Vietnamese. *Journal of Phonetics*, *37*(1), 79–96. https://doi.org/10.1016/j.wocn.2008.09.003

Brunelle, M., Nguyên, D. D., & Nguyên, K. H. (2010). A Laryngographic and Laryngoscopic Study of Northern Vietnamese Tones. *Phonetica*, *67*(3), 147–169. https://doi.org/10.1159/000321053

Bundgaard-Nielsen, R. L., Best, C. T., Kroos, C., & Tyler, M. D. (2012). Second language learners' vocabulary expansion is associated with improved second language vowel intelligibility. *Applied Psycholinguistics*, *33*(3), 643–664. https://doi.org/10.1017/S0142716411000518

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011a). Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese

vowel categories. *Applied Psycholinguistics*, *32*(1), 51–67. https://doi.org/10.1017/S0142716410000287

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011b). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition*, *33*(3), 433–461. https://doi.org/10.1017/S0272263111000040

Burnham, D., Kuratate, T., McBride-Chang, C., & Mattock, K. (2009). *Making speech three-dimensional: Adding tone to consonant- and vowel-based speech perception and language acquisition research, quantification and theory*. http://purl.org/au-research/grants/arc/DP0988201

Burnham, D., Kasisopa, B., Reid, A., Luksaneeyanawin, S., Lacerda, F., Attina, V., Rattanasone, N. X., Schwarz, I.-C., & Webster, D. (2015). Universality and language-specific experience in the perception of lexical tone and pitch. *Applied Psycholinguistics*, *36*(6), 1459–1491. https://doi.org/10.1017/S0142716414000496

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.

Chao. Y.R. (1930). A system of tone-letters. *Le Maitre Phonetique*, *45*, 24–27.

Chen, J., Best, C. T., & Antoniou, M. (2019). Cognitive Factors in Thai-Naïve Mandarin Speakers' Imitation of Thai Lexical Tones. *Proc. Interspeech 2019*, 2653–2657. https://doi.org/10.21437/Interspeech.2019-1403

Chen, J., Best, C. T., Antoniou, M., & Kasisopa, B. (2018). *Mapping and comparing East and Southeast Asian language tones*. Australia Linguistic Society annual conference, Adelaide.

Chen, J., Best, C. T., Antoniou, M., & Kasisopa, B. (2019). Cognitive factors in perception of Thai tones by naïve Mandarin listeners. In S. Calhoun, P. Escudero, M. Tabain, & P.

Warren (Eds.), *Proceedings of the 19th ICPhS,* (pp. 1684–1688). Australasian Speech Science and Technology Association Inc.

Chen, J., Best, C. T., & Antoniou, M. (2020). Native phonological and phonetic influences in perceptual assimilation of monosyllabic Thai lexical tones by Mandarin and Vietnamese listeners. *Journal of Phonetics*, *83*, 101013. https://doi.org/10.1016/j.wocn.2020.101013

Chiao, W.-S., Kabak, B., & Braun, B. (2011). When more is less: Non-native perception of level tone contrasts. *Proceedings of the Psycholinguistic Representation of Tone Conference*, 42-45.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* Harper & Row, Publishers.

Chuang, C. -K., & Hiki, S. (1972). Acoustical Features and Perceptual Cues of the Four Tones of Standard Colloquial Chinese. *The Journal of the Acoustical Society of America*, *52*(1), 146–146. https://doi.org/10.1121/1.1981919

Clements, G. N., Michaud, A., & Patin, C. (2011). Do we need tone features. In J. A. Goldsmith, E. Hume, & L. Wetzels, *Tones and Features: Phonetic and Phonological Perspectives*. Berlin: De Gruyter Mouton.

Davis, K., & Kuhl, P. K. (1994). Tests of the perceptual magnet effect for American English /k/ and /g/. *The Journal of the Acoustical Society of America*, *95*(5), 2976–2976. https://doi.org/10.1121/1.408982

Diehl, R. L., & Kluender, K. R. (1989). On the Objects of Speech Perception. *Ecological Psychology*, *1*(2), 121–144. https://doi.org/10.1207/s15326969eco0102_2

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology*, *55*(1), 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Duanmu, S. (1990). *A formal study of syllable, tone, stress and domain in Chinese languages*. Massachusetts Institute of Technology.

Duanmu, S. (1994). Against Contour Tone Units. *Linguistic Inquiry*, *25*(4), 555–608.

Elvin, J., Escudero, P., & Vasiliev, P. (2014). Spanish is better than English for discriminating Portuguese vowels: Acoustic similarity versus vowel inventory size. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01188

Erickson, D. (1976). *A Physiological Analysis of the Tones of Thai.* [Doctoral dissertation, University of Connecticut]. https://search.proquest.com/docview/302781298/citation/77296413DF0D40C9PQ/1

Erickson, D., Liberman, M. Y., & Niimi, S. (1976). The geniohyoid and the role of the strap muscles in pitch control. *The Journal of the Acoustical Society of America*, *60*(S1), S63–S63. https://doi.org/10.1121/1.2003454

Erickson, Donna, & Abramson, A. S. (2013). F0, EMG and Tonogenesis in Thai. *Journal of Nagoya Gakuin University (Language and Culture): Collected Papers in Honor of Prof. Katsumasa Shimizu*, 24–1.

Escudero, P. (2005). *Linguistic Perception and Second Language Acquisition: Explaining the Attainment of Optimal Phonological Categorization*. LOT.

Escudero, P. (2009). Linguistic perception of "similar" L2 sounds. In P. Boersma & S. Hamann (Eds.), *Phonology in perception* (Vol. 15, pp. 152–190). Mouton de Gruyter.

Escudero, P., & Vasiliev, P. (2011). Cross-language acoustic similarity predicts perceptual assimilation of Canadian English and Canadian French vowels. *The Journal of the Acoustical Society of America*, *130*(5), EL277–283. https://doi.org/10.1121/1.3632043

Escudero, P., & Williams, D. (2011). Perceptual assimilation of Dutch vowels by Peruvian Spanish listeners. *The Journal of the Acoustical Society of America*, *129*(1), EL1-7. https://doi.org/10.1121/1.3525042

Ewan, W. G. (1975). Explaining the intrinsic pitch of vowels. *The Journal of the Acoustical Society of America*, *58*(S1), S40. https://doi.org/10.1121/1.2002115

Faris, M. M., Best, C. T., & Tyler, M. D. (2016). An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized. *The Journal of the Acoustical Society of America*, *139*(1), EL1-5. https://doi.org/10.1121/1.4939608

Faris, M. M., Best, C. T., & Tyler, M. D. (2018). Discrimination of uncategorised non-native vowel contrasts is modulated by perceived overlap with native phonological categories. *Journal of Phonetics*, *70*, 1–19. https://doi.org/10.1016/j.wocn.2018.05.003

Flege, J. E. (1987). The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*, 47–65.

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.

Flege, J. E., Schirru, C., & MacKay, I. R. A. (2003). Interaction between the native and second language phonetic subsystems. *Speech Communication*, *40*(4), 467–491. https://doi.org/10.1016/S0167-6393(02)00128-0

Flege, J. E. (1992). The intelligibility of English vowels spoken by British and Dutch talkers. In R. D. Kent (Ed.), *Intelligibility in speech disorders: Theory, measurement, and management* (pp. 157–232).

Flege, J. E. (1991). Perception and production: The relevance of phonetic input to L2 phonological learning. In T. Huebner & C. A. Ferguson (Eds.), *Crosscurrents in second language acquisition and linguistic theories* (pp. 249–289).

Flege, J. E., & Fletcher, K. L. (1992). Talker and listener effects on degree of perceived foreign accent. *The Journal of the Acoustical Society of America*, *91*(1), 370–389. https://doi.org/10.1121/1.402780

Flege, J. E., & Hillenbrand, J. (1984). Limits on phonetic accuracy in foreign language speech production. *The Journal of the Acoustical Society of America*, *76*(3), 708–721. https://doi.org/10.1121/1.391257

Flege, J. E., & MacKay, I. R. A. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition*, *26*(1), 1–34. https://doi.org/10.1017/S0272263104026117

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese Adults can Learn to Produce English /I/ and /l/ Accurately. *Language and Speech*, *38*(1), 25–55. https://doi.org/10.1177/002383099503800102

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). York Press.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*(1), 3–28.

Fowler, C. A. (1989). Real Objects of Speech Perception: A Commentary on Diehl and Kluender. *Ecological Psychology*, *1*(2), 145–160. https://doi.org/10.1207/s15326969eco0102_3

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (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. https://doi.org/10.1016/S0749-596X(03)00072-X

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fujisaki, H., & Kawashima, T. (1970). *Some experiments on speech perception and a model for the perceptual mechanism.* (pp. 29, 207–214) [Annual Report of the Engineering Research Institute]. University of Tokyo.

Gandour, J. T. (1978). The perception of tone. In V. A. Fromkin (Ed.), *Tone: A linguistic survey* (pp. 41–76). Academic Press.

Gandour, J. T., & Harshman, R. A. (1978). Crosslanguage differences in tone perception: A multidimensional scaling investigation. *Language and Speech*, *21*(1), 1–33.

Gao, M. (2009). Gestural coordination among vowel, consonant and tone gestures in Mandarin Chinese. *Chinese Journal of Phonetics*, *2*. http://urn.kb.se/resolve?urn=urn:nbn:se:du-5137

Gottfried, T. L., Staby, A. M., & Ziemer, C. J. (2004). Musical experience and Mandarin tone discrimination and imitation. *The Journal of the Acoustical Society of America*, *115*(5), 2545–2545. https://doi.org/10.1121/1.4783674

Gruber, J. (1964). *The distinctive features of tone*. [Unpublished manuscript]

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm *R* Package. *Journal of Statistical Software*, *33*(2). https://doi.org/10.18637/jss.v033.i02

Halekoh, U., & Hojsgaard, S. (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest. *Journal of Statistical Software*, *59*(9), 1–30.

Hallé, P. A., Best, C. T., & Levitt, A. (1999). Phonetic vs. Phonological influences on French listeners' perception of American English approximants. *Journal of Phonetics*, *27*(3), 281–306. https://doi.org/10.1006/jpho.1999.0097

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

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279. https://doi.org/10.1016/j.wocn.2011.11.001

Hao, Y.-C. (2017). 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. https://doi.org/10.1016/j.wocn.2015.10.003

Haudricourt, A.-G. (1954). De l'origine des tons en vietnamien. *Journal Asiatique*, *142*, 69–82.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Howie, J. M. (1972). Some experiments on the perception of Mandarin tones. *Proceedings of the 7th International Congress of Phonetic Sciences*, 900–904.

Hsieh, L., & Yu, Y. (2006). Tone sandhi effect on Chinese speech perception. *The Journal of the Acoustical Society of America*, *120*(5), 3086–3087. https://doi.org/10.1121/1.4787460

Huang, T., & Johnson, K. (2010). Language specificity in speech perception: Perception of Mandarin tones by native and nonnative listeners. *Phonetica*, *67*(4), 243–267. https://doi.org/10.1159/000327392

Hyman, L. M. (2006). Word-prosodic typology. *Phonology*, *23*(2), 225–257. https://doi.org/10.1017/S0952675706000893

Hyman, L. M. (2011). Do tones have features? In J. A. Goldsmith, E. Hume, & L. Wetzels (Eds.), *Tones and Features*. De Gruyter. https://doi.org/10.1515/9783110246223.50

Hyman, L. M., & VanBik, K. (2004). Directional Rule Application and Output Problems in Hakha Lai Tone. *Language and Linguistics*, *5*(4), 821–861.

Hyman, L., & VanBik, K. (2002). *Tone and syllable structure in Hakha-Lai*. Annual Meeting of the Berkeley Linguistics Society.

Iverson, P., & Kuhl, P. K. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *The Journal of the Acoustical Society of America*, *99*(2), 1130–1140. https://doi.org/10.1121/1.415234

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

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2

exposure. *The Journal of the Acoustical Society of America*, *119*(2), 1118–1130. https://doi.org/10.1121/1.2151806

Jusczyk, P. (1997). *The Discovery of Spoken Language*. MIT Press.

Kang, Y. (2014). Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: A corpus study. *Journal of Phonetics*, *45*, 76–90. https://doi.org/10.1016/j.wocn.2014.03.005

Kang, Y., & Han, S. (2013). Tonogenesis in early Contemporary Seoul Korean: A longitudinal case study. *Lingua*, *134*, 62–74. https://doi.org/10.1016/j.lingua.2013.06.002

Kirby, J. (2010). Dialect experience in Vietnamese tone perception. *The Journal of the Acoustical Society of America*, *127*(6), 3749–3757. https://doi.org/10.1121/1.3327793

Kiriloff, C. (1969). On the Auditory Perception of Tones in Mandarin. *Phonetica*, *20*, 63–67. https://doi.org/10.1159/000259274

Krebs-Lazendic, L., & Best, C. (2013). First language suprasegmentally-conditioned syllable length distinctions influence perception and production of second language vowel contrasts. *Laboratory Phonology*, *4*(2), 435–474.

Kuang, J. (2013). The Tonal Space of Contrastive Five Level Tones. *Phonetica*, *70*(1–2), 1–23. https://doi.org/10.1159/000353853

Kuhl, P. K., & Iverson, P. (1995). Linguistic experiencce and the "Perceptual Magnet Effect." In W. Strange, *Speech perception and linguistic experience: Issues in cross-language research* (pp. 121–154). York Press.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, *190*(4209), 69–72. https://doi.org/10.1126/science.1166301

Kuhl, P. 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. https://doi.org/10.3758/BF03212211

Kuhl, P. K. (1993). Innate Predispositions and the Effects of Experience in Speech Perception: The Native Language Magnet Theory. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage, & J. Morton (Eds.), *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp. 259–274). Springer Netherlands. https://doi.org/10.1007/978-94-015-8234-6_22

Kuhl, P. K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology*, *4*(6), 812–822. https://doi.org/10.1016/0959-4388(94)90128-7

Lado, R. (1957). *Linguistics across cultures*. University of Michigan Press.

Lee, C.-Y., Tao, L., & Bond, Z. S. (2009). Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *Journal of Phonetics*, *37*(1), 1–15. https://doi.org/10.1016/j.wocn.2008.08.001

Lee, Y. S., Vakoch, D. A., & Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research*, *25*(5), 527–542.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. https://doi.org/10.18637/

Levy, E. S. (2009). On the assimilation-discrimination relationship in American English adults' French vowel learning. *The Journal of the Acoustical Society of America*, *126*(5), 2670–2682. https://doi.org/10.1121/1.3224715

Li, F. K. (1977). *A handbook of comparative Tai*. University of Hawai'i Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

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*. https://doi.org/10.1037/h0044417

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

Lin, Y.-H. (1989). *Autosegmental treatment of segmental processes in Chinese phonology* [Unpublished Doctoral dissertation]. University of Texas at Austin.

Liu, S., & Samuel, A. G. (2004). Perception of Mandarin Lexical Tones when F0 Information is Neutralized. *Language and Speech*, *47*(2), 109–138. https://doi.org/10.1177/00238309040470020101

Llompart, M., & Reinisch, E. (2018). Imitation in a Second Language Relies on Phonological Categories but Does Not Reflect the Productive Usage of Difficult Sound Contrasts. *Language and Speech*, *62*(3), 594–622. https://doi.org/10.1177/0023830918803978

Lobanov, B. M. (1971). Classification of Russian Vowels Spoken by Different Speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606–608. https://doi.org/10.1121/1.1912396

Logan, G. D. (1979). On the use of a concurrent memory load to measure attention and automaticity. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(2), 189–207. https://doi.org/10.1037/0096-1523.5.2.189

MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, *2*(4), 369–390. https://doi.org/10.1017/S0142716400009796

MacKay, I. R. A., Meador, D., & Flege, J. E. (2001). The Identification of English Consonants by Native Speakers of Italian. *Phonetica*, *58*(1–2), 103–125. https://doi.org/10.1159/000028490

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Psychology press.

Maddieson, I. (2013). Tone. In M. Haspelmath, M. Dryer, D. Gil, & B. Comrie (Eds.), *The world atlas of language structures online*. Max Planck Institute for Evolutionary Anthropology Leipzig.

Maddieson, I. (1976). A further note on tone and consonants. *UCLA Working Papers in Phonetics*, *33*, 131–159.

Maddieson, I. (1972). Tone system typology and distinctive features. *Proceedings of the 7th Internation Congress of Phonetic Science*, 958–961.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409. https://doi.org/10.1037/0096-1523.33.2.391

Magnuson, J. S., & Yamada, R. A. (1994). Talker variability and the identification of American English /r/ and /l/ by Japanese subjects. *The Journal of the Acoustical Society of America*, *95*(5), 2872–2872. https://doi.org/10.1121/1.409439

Maspero, H. (1912). Etudes sur la phonétique historique de la langue annamite. Les initiales. *Bulletin de l'École Française d'Extrême-Orient*, *12*(1), 1–124.

Massaro, D. W., & Oden, G. C. (1980). Evaluation and integration of acoustic features in speech perception. *The Journal of the Acoustical Society of America*, *67*(3), 996–1013. https://doi.org/10.1121/1.383941

Mei, T. (1970). Tones and Prosody in Middle Chinese and The Origin of The Rising Tone. *Harvard Journal of Asiatic Studies*, *30*, 86–110. JSTOR. https://doi.org/10.2307/2718766

Miller, J. (1996). The sampling distribution of d′. *Perception & Psychophysics*, *58*(1), 65–72. https://doi.org/10.3758/BF03205476

Miller, J. D., Wier, C. C., Pastore, R. E., Kelly, W. J., & Dooling, R. J. (1976). Discrimination and labeling of noise–buzz sequences with varying noise-lead times: An example of categorical perception. *The Journal of the Acoustical Society of America*, *60*(2), 410–417. https://doi.org/10.1121/1.381097

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, *18*(5), 331–340. https://doi.org/10.3758/BF03211209

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390. https://doi.org/10.3758/BF03210878

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378. https://doi.org/10.1121/1.397688

Nespor, M., & Vogel, I. (2007). *Prosodic phonology: With a new foreword*. Walter de Gruyter.

Nhàn, N. T. (1984). *The syllabeme and patterns of word formation in vietnamese* [PhD Dessertation]. New York University.

Nguyen, V. L., & Edmondson, J. A. (1998). Tones and voice quality in modern northern Vietnamese: Instrumental case studies. *Mon-Khmer Studies*, *28*, 1–18.

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: Eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language*, *90*, 103–125. https://doi.org/10.1016/j.jml.2016.03.005

Nusbaum, H., & Morin, T. M. (1992). Paying Attention to Differences Among Talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech perception, production and linguistic structure* (pp. 113–134). Ohmasha Publishing.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1718–1725. https://doi.org/10.1121/1.414696

Pham, A. H. (2004). *Vietnamese tone: A new analysis*. Routledge.

Pham, A. H. (2003). *The key phonetic properties of Vietnamese tone: A reassessment*. 1703–1706.

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

Polka, L. (1992). Characterizing the influence of native language experience on adult speech perception. *Perception & Psychophysics*, *52*(1), 37–52. https://doi.org/10.3758/BF03206758

Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *The Journal of the Acoustical Society of America*, *97*(2), 1286–1296.

R Core Team. (2018). *R: A language and environment for statistical computing* [Manual]. https://www.R-project.org/

Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Rattanasone, N. X., & Best, C. (2015). Perceptual assimilation of lexical tone: The roles of language experience and visual information. *Attention Perception & Psychophysics*, *77*(2), 571–591. https://doi.org/10.3758/s13414-014-0791-3

Remijsen, B. (2016). Tone. In *Oxford Research Encyclopedia of Linguistics*.

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. Scopus. https://doi.org/10.1016/0167-6393(85)90039-1

Rojczyk, A. (2012a). *Phonetic and phonological mode in second-language speech: VOT imitation.* EUROSLA 2012, Poznań Poland.

Rojczyk, A. (2012b). Phonetic imitation of L2 vowels in a rapid shadowing task. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference* (pp. 66–76).

Rojczyk, A., Porzuczek, A., & Bergier, M. (2013). Immediate and Distracted Imitation in Second-Language Speech: Unreleased Plosives in English. *Research in Language*, *11*(1), 3–18. https://doi.org/10.2478/v10015-012-0007-7

Sagart, L., Hallé, P. A., Boysson-Bardies, B. de, & Arabia-Guidet, C. (1986). Tone production in modern standard chinese: An electromyographic investigation. *Cahiers de Linguistique - Asie Orientale*, *15*(2), 205–221. https://doi.org/10.3406/clao.1986.1204

Sampson, G. (1969). A Note on Wang's "Phonological Features of Tone." *International Journal of American Linguistics*, *35*(1), 62–66. https://doi.org/10.1086/465041

Schwanhausser, B., & Burnham, D. (2005). Lexical Tone and Pitch Perception in Tone and Non-Tone Language Speakers. *INTERSPEECH 2005*, 1701–1704.

Selkirk, E., & Shen, T. (1990). Prosodic domains in Shanghai Chinese. In S. Inkelas & D. Zec (Eds.), *The Phonology-Syntax Connection* (pp. 313–338). University of Chicago Press.

Shaw, J. A., Chen, W., Proctor, M. I., & Derrick, D. (2016). Influences of Tone on Vowel Articulation in Mandarin Chinese. *Journal of Speech, Language, and Hearing Research*, *59*(6), S1566–S1574. https://doi.org/10.1044/2015_JSLHR-S-15-0031

Shaw, J. A., & Tyler, M. D. (2020). Effects of vowel coproduction on the time course of tone recognition. *The Journal of the Acoustical Society of America*, *147*(4), 2511–2524. https://doi.org/10.1121/10.0001103

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(3), 243–261. Scopus. https://doi.org/10.1017/S0142716400001417

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429. https://doi.org/10.3758/BF03194890

Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, *23*(2), 287–308. https://doi.org/10.1017/S0952675706000911

Smolensky, P., & Prince, A. (1993). Optimality Theory: Constraint interaction in generative grammar. *Optimality Theory in Phonology*, 3.

Snodgrass, J., Levy-Berger, G., & Haydon, M. (1985). *Human experimental psychology*. Oxford University Press.

So, C. K. (2012). Cross-language categorization of monosyllabic foreign tones: Effects of phonological and phonetic properties of native language. In T. Stolz, N. Nau, & Stroh (Eds.), *Monosyllables: From Phonology to Typology* (pp. 55–69).

So, C. K., & Best, C. T. (2010a). Discrimination and categorization of Mandarin tones by Cantonese speakers: The role of native phonological and phonetic properties. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*.

So, C. K., & Best, C. T. (2010b). 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*(1), 133–145. 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

Stevens, K. N., Keyser, S. J., & Kawasaki, H. (1986). Toward a phonetic and phonological theory of redundant features. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and Variability in Speech Processes* (pp. 426–449). Taylor & Francis.

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

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., & Nishi, K. (2001). Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *The Journal of the Acoustical Society of America*, *109*(4), 1691–1704.

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners. In J. G. H. Edwards & M. L. Zampini, *Phonology and second language acquisition*.

Studdert-Kennedy, M., & Shankweiler, D. (1970). Hemispheric Specialization for Speech Perception. *The Journal of the Acoustical Society of America*, *48*(2B), 579–594. https://doi.org/10.1121/1.1912174

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, *21*(2), 241–299. https://doi.org/10.1017/S0140525X98001174

Sussman, J. E., & Lauckner-Morano, V. J. (1995). Further tests of the '"perceptual magnet effect"' in the perception of [i]: Identification and change/no-change discrimination. *The Journal of the Acoustical Society of America*, *97*(1), 539–552. https://doi.org/10.1121/1.413111

Svantesson, J. O. (1989). Tonogenetic mechanisms in northern Mon-Khmer. *Phonetica*, *46*(1–3), 60–79. https://doi.org/10.1159/000261829

The International Phonetic Alphabet. (2015). IPA Chart, http://www.internationalphoneticassociation.org/content/ipa-chart, available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright © 2015 International Phonetic Association.

Thurgood, G. (2002). Vietnamese and tonogenesis: Revising the model and the analysis. *Diachronica*, *19*(2), 333–363. https://doi.org/10.1075/dia.19.2.04thu

Thurgood, G. (2007). Tonogenesis revisited: Revising the model and the analysis. *Studies in Tai and Southeast Asian Linguistics*, 263–291.

Tingsabadh, M. R. K. (2001). Thai tone geography. In M. R. Kalaya Tingsabadh & A. S. Abramson (Eds.), *Essays in Tai Linguistics* (pp. 205–228). Chulalongkorn University Press.

Tsukada, K. (2019). Are Asian Language Speakers Similar or Different? The Perception of Mandarin Lexical Tones by Naïve Listeners from Tonal Language Backgrounds: A Preliminary Comparison of Thai and Vietnamese Listeners. *Australian Journal of Linguistics*, *0*(0), 1–18. https://doi.org/10.1080/07268602.2019.1620681

Tsukada, K., & Han, J.-I. (2019). The perception of Mandarin lexical tones by native Korean speakers differing in their experience with Mandarin. *Second Language Research*, *35*(3), 305–318. https://doi.org/10.1177/0267658318775155

Tsukada, K., & Kondo, M. (2018). The Perception of Mandarin Lexical Tones by Native Speakers of Burmese. *Language and Speech*, *62*(4), 625–640. https://doi.org/10.1177/0023830918806550

Tsukada, K., Kondo, M., & Sunaoka, K. (2016). The perception of Mandarin lexical tones by native Japanese adult listeners with and without Mandarin learning experience. *Journal of Second Language Pronunciation*, *2*(2), 225–252. https://doi.org/10.1075/jslp.2.2.05tsu

Tyler, M. D., Best, C. T., Faber, A., & Levitt, A. G. (2014). Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica*, *71*(1), 4–21. https://doi.org/10.1159/000356237

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

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (4th ed.). Springer. http://www.stats.ox.ac.uk/pub/MASS4/

Vũ, T. P. (1981). *The acoustic and perceptual nature of tone in Vietnamese* [PhD Dessertation, Australia National University]. https://openresearch-repository.anu.edu.au/handle/1885/12396

Wang, W. S.-Y. (1967). Phonological Features of Tone. *International Journal of American Linguistics*, *33*(2), 93–105. https://doi.org/10.1086/464946

Wang, X. (2013). Perception of Mandarin Tones: The Effect of L1 Background and Training. *The Modern Language Journal*, *97*(1), 144–160. https://doi.org/10.1111/j.1540-4781.2013.01386.x

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*(6), 3649–3658. https://doi.org/10.1121/1.428217

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

Wayland, R. (1997). Non-native Production of Thai: Acoustic Measurements and Accentedness Ratings. *Applied Linguistics*, *18*(3), 345–373. https://doi.org/10.1093/applin/18.3.345

Wayland, R., & Guion, S. (2005). Sound changes following the loss of /r/ in Khmer: A new tonogenetic mechanism? *Mon-Khmer Studies*, *35*, 55–82.

Werker, J. F., & Tees, R. C. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *The Journal of the Acoustical Society of America*, *75*(6), 1866–1878.

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception & Psychophysics*, *37*(1), 35–44. https://doi.org/10.3758/BF03207136

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics*, *23*(3), 349–366. https://doi.org/10.1016/S0095-4470(95)80165-0

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S. N., & Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, *59*, 122–143. https://doi.org/10.1016/j.wocn.2016.09.004

Wiener, S., & Lee, C.-Y. (2020). Multi-Talker Speech Promotes Greater Knowledge-Based Spoken Mandarin Word Recognition in First and Second Language Listeners. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00214

Wong, P. C. M., & Diehl, Randy. L. (2003). Perceptual normalization for inter-and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, *46*(2), 413–421.

Woo, N. (1969). *Prosody and Phonology* [PhD Dessertation]. MIT.

Wu, M., Bundgaard-Nielsen, R. L., Baker, B., Best, C. T., & Fletcher, J. (2015). Perception of Cantonese tones by Mandarin speakers. *ICPhS*.

Wu, X., Munro, M. J., & Wang, Y. (2014). Tone assimilation by Mandarin and Thai listeners with and without L2 experience. *Journal of Phonetics*, *46*, 86–100. https://doi.org/10.1016/j.wocn.2014.06.005

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

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics*, *52*(4), 376–392. https://doi.org/10.3758/BF03206698

Yeh, C.-H., & Lin, Y.-H. (2012). *The Effect of Tone Sandhi on Speech Perception of Taiwanese Falling Tones*. 1–4.

Yip, M. (2001). Tonal features, tonal inventories and phonetic targets. *UCL Working Papers in Linguistics*, *13*, 161–188.

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

Yu, Y. H., Shafer, V. L., & Sussman, E. S. (2017). Neurophysiological and Behavioral Responses of Mandarin Lexical Tone Processing. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00095

Zheng, Q. (2014). Effects of Vowels on Mandarin Tone Categorical Perception. *Acta Psychologica Sinica*, *46*(9), 1223–1231. https://doi.org/10.3724/SP.J.1041.2014.01223

Appendix A Supplementary materials for Chapter 5

Acoustic measures of tones in Thai, Mandarin, Northern Vietnamese and Southern Vietnamese

While normalised F0 contours offer qualitative visual comparisons of temporally dynamic characteristics of tones across languages, discrete features can allow quantitative comparisons, including duration (which was removed via time-normalisation for Figure 1). Table A presents a summary of six acoustic measures for each tone per each language. According to a previous multidimensional scaling study, duration, and F0 mean, direction, extreme endpoint, and slope have been found to correlate with perception of tones by native listeners of Thai and Yoruba (Gandour, 1978). To capture these features, we calculated syllable duration, $F0_{onset}$, $F0_{offset}$, $F0_{mean}$, and $F0_{excursion}$ (maximum to minimum), which had been used to characterise level tone contrasts in a previous study (Kuang, 2013), and we added one more measure, $F0_{max\_location}$ ratio, (i.e., relative location of the F0 peak as a proportion of the duration of the tone) to distinguish differently-timed peaks in convex and concave contours (e.g. T241 and T315).

Ideally it would be helpful to provide statistical comparisons based on these discrete measures by individual linear mixed-effect models, or principle component analysis or machine learning classification algorithms. We have conducted all of these on our acoustic data but each was accompanied by serious methodologically inherent problems as pointed out by reviewers. Significant differences in acoustic dimensions as shown in the linear mixed effects models and the Tukey adjusted multiple comparisons do not necessarily lead to problems in perceptual assimilations, which depend on listeners' relative weighting of the individual acoustic feature in their native language. Principle component analysis offers visual comparisons, but its dimension reduction algorithm does not necessarily capture relative weightings in the same way as human

listeners. Other machine learning classification algorithms, such as Support Vector Machine, could be applied but the interpretation would have to be treated very cautiously given our relatively small acoustic data set. Due to these issues and possible distractions of these models, we do not include the model-based comparisons we conducted, but we provide the measurements we took in the table below.

Table A.1 Acoustic measures of tones in Thai (20 tokens per tone), Mandarin, NV and SV (32 tokens per tone). All measures, $F0_{mean}$, $F0_{onset}$, $F0_{off}$, $F0_{excursion,}$ are Lobanov-normalised (Lobanov, 1971). * indicates that three decimal places were kept to show the real value was not equal to zero.

	Duration (ms)		$F0_{mean}$		$F0_{onset}$		$F0_{offset}$		$F0_{excursion}$		$F0_{max\_location}$ (%)	
Tones	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
T45	619	70	0.05	0.05	0.02	0.04	0.17	0.08	0.17	0.05	100	2
T33	640	85	-0.01	0.08	0.01	0.04	-0.08	0.09	0.11	0.07	34	23
T21	622	75	-0.1	0.04	-0.01	0.05	-0.21	0.1	0.22	0.1	22	7
T315	642	82	-0.08	0.05	-0.03	0.04	0.01	0.12	0.19	0.09	72	41
T241	565	101	0.13	0.06	0.09	0.05	-0.02	0.09	0.21	0.1	53	14
M55	663	137	0.14	0.04	0.08	0.07	0.15	0.05	0.09	0.05	74	27
M35	613	106	-0.05	0.05	-0.11	0.04	0.14	0.13	0.3	0.08	99	3
M214	745	137	-0.23	0.1	-0.12	0.08	-0.18	0.09	0.43	0.26	42	32
M51	506	127	0.08	0.06	0.19	0.14	-0.11	0.18	0.44	0.19	29	13
NV44	435	55	0.09	0.07	0.04	0.04	0.09	0.06	0.1	0.07	62	17
NV22	501	54	-0.1	0.08	-0.07	0.08	-0.16	0.09	0.11	0.06	25	12
NV35	431	61	*0.001	0.06	-0.05	0.04	0.23	0.14	0.31	0.13	99	3
NV21	286	50	-0.01	0.05	*0.004	0.04	-0.08	0.09	0.12	0.08	51	27
NV415	402	56	0.07	0.11	0.01	0.06	0.3	0.17	0.43	0.18	97	13
NV214	406	63	-0.13	0.06	-0.04	0.06	-0.22	0.07	0.21	0.07	18	9
SV44	469	37	0.08	0.04	0.08	0.03	-0.01	0.13	0.16	0.11	58	28
SV22	501	41	-0.11	0.05	-0.04	0.07	-0.2	0.07	0.17	0.06	16	8
SV35	465	41	0.19	0.05	0.05	0.05	0.48	0.12	0.44	0.1	100	2
SV21	456	40	-0.17	0.07	-0.1	0.1	-0.11	0.19	0.26	0.23	49	43
SV415	493	52	-0.07	0.06	-0.09	0.08	0.24	0.1	0.48	0.15	100	0
SV214	483	47	-0.1	0.06	-0.09	0.09	0.21	0.1	0.49	0.17	100	0

GAMM modeling for comparing Northern and Southern Vietnamese tone merger.

The tone merger of SV214 (*hỏi*) and SV415 (*ngã*) in Southern Vietnamese makes the phonological system of Southern Vietnamese different from that of Northern Vietnamese. Given the importance of this difference and its potential influences on perceptual assimilation patterns of Thai tones by these two groups, we verified the tone merger in Southern Vietnamese in terms of the F0 contours of these two tones in our acoustic study. In order to produce a formally holistic and dynamic comparison of F0 contours, we employed a General Additive Mixed Model (GAMM), which is a non-linear regression method that does not require aggregation or the pre-selection of a fixed time point in the contours. The method can detect general patterns over dynamically varying data, while at the same time accounting for subject and item-related variability (Nixon et al., 2016; Wieling, 2018). In this way, it can uncover patterns that are obscured when data are aggregated or when a single time point is arbitrarily chosen.

Our dependent variable was the Lobanov-normalised F0 values which we modelled by using *smooths*. These smooths model non-linear patterns by combining a pre-specified number of basis functions. First, we built one model for each language/dialect separately. In each model, we set up smooths, *s(Time, by = tone),* over time separately for each tone, *hỏi* and *ngã.* In order to take speaker variation into consideration, we also modelled a non-linear random effect of speaker for each tone type via factor smoothing functions in the R package *mgcv*, as suggested by previous studies on articulatory data (Wieling, 2018; Wieling et al., 2016). *s(Time, subject, by = tone, bs = "fs", m = 1)* is a factor smoothing function that models a non-linear difference over *Time* with respect to general time pattern for each *subject* (equal to random factors in the linear mixed-effect model). After we fitted the model, we used the *acf_resid* function in the *itsadug* package in R to obtain the autocorrelation measure and fed the autocorrelation measures at lag 1 to the *rho setting*

in the new model. Figure B1 shows the non-linear smooths of the tone contrasts in each Vietnamese regional dialect separately.

Visual comparison of the two panels in Figure B1 suggests that the contrast exists in Northern Vietnamese but not in Southern Vietnamese. In order to formally compare the contrast in two Vietnamese dialects, we employed two approaches as suggested by (Wieling, 2018). The first approach is to refit the model with a binary difference smooth, which models the difference of the contrasts. In Northern Vietnamese, the difference is significant $F(5.11) = 18.66$, $p < .001$, but in Southern Vietnamese, the different is not significant, $F(2.00) = 0.16$, $p = .86$. The second approach is to refit the model with an ordered factor difference smooth. The results also shows that the contrast is significantly different for Northern Vietnamese, $F(5.23) = 9.99$, $p < .001$, but not Southern Vietnamese, $F(1.01) = 0.10$, $p = .76$.

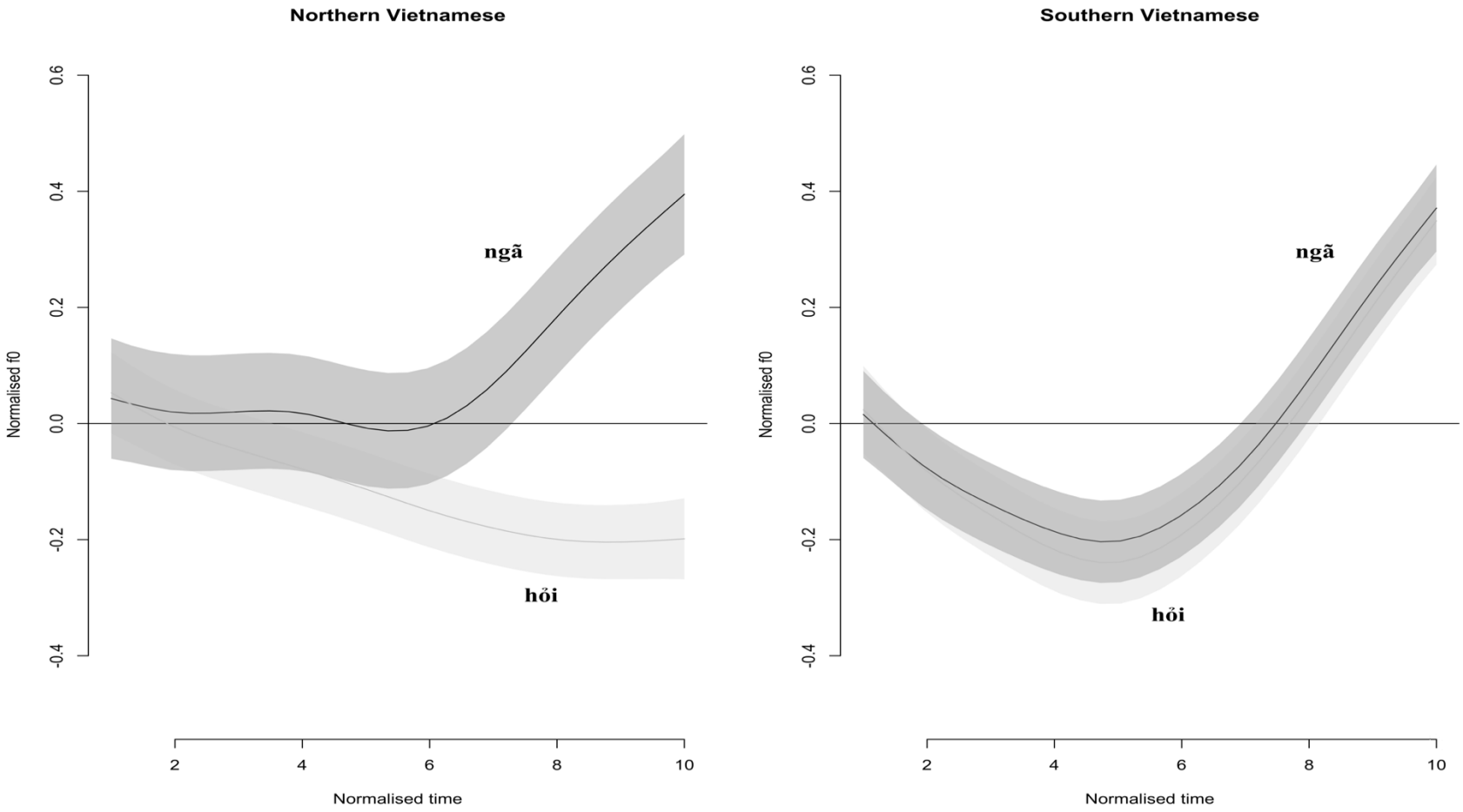


Figure A.1 Non-linear smooths for *hỏi* and *ngã* in Northern Vietnamese (left) and Southern Vietnamese (right). The pointwise 95%-confidence intervals are shown by ribbons.

Table A.2 Mean percentage of choice (%), mean category-goodness ratings and mean response times (RT, ms) for categorisations of each Thai tone to the tones in each listener language. "-" means no response. Ratings: 1 = poor, 7 = perfect.

Native categories	T45			T33			T21			T315			T241		
	%	rating	RT	%	rating	RT	%	rating	RT	%	rating	RT	%	rating	RT
M55	0.7	2	1869	92.5	4.8	800	11.9	4.2	1376	-	-	-	48.1	4.3	1090
M35	88.4	5	711	0.4	5	1244	1.1	4	1996	79.6	5.3	667	-	-	-
M214	10.9	4.4	1414	2.5	3.5	1250	59.1	4.9	947	20.4	4.4	877	1.1	4	1582
M51	-	-	-	4.6	3.5	1533	27.9	4.6	1053	-	-	-	50.8	4.3	860
NV44	5.9	2.6	1393	29.4	4.9	821	1	1.3	307	0.7	3.5	334	92.3	5.3	620
NV22	1	2	1628	69.2	5	704	77.4	4.9	799	0.3	3	982	4.3	5	1215
NV35	55.1	4.1	924	-	-	-	0.3	6	1127	51	4.7	800	0.4	1	1630
NV21	3.5	3.5	1290	1	2	1112	2.1	2.7	911	2.1	2.4	1263	1.9	1.6	2126
NV415	22.5	3.3	1440	0.3	1	1795	-	-	-	26.5	3.6	1207	0.4	3	2449
NV214	11.9	4.2	1596	-	-	-	19.1	3.9	1089	19.4	3.5	1054	0.7	1	2178
SV44	6.3	3.2	1030	35.3	5.2	875	6.2	4	1042	4.5	2.5	1072	88.9	5.3	743
SV22	0.3	5	1758	60.6	4.8	835	83.7	5	641	-	-	-	4.9	4.8	1003
SV35	31.1	4.5	1086	-	-	-	0.3	7	604	9.2	4.6	669	2.1	4.5	928
SV21	31	4.7	1541	3.5	4.4	1175	5.6	3.4	1102	0.7	4.5	499	2.8	4.2	717
SV214	31.2	3.8	1139	0.7	2	1617	4.2	5	553	85.6	4.8	1013	1.4	4.5	718

Statistical tests for determining Categorised and UnCategorised assimilations.

Table A.3 Testing native category choices against chance level (significant results are in bold, p < .05)

Thai targets	Native tone choices	*t*	*df*	*p*
T21	M214	4.13	9	**0.001**
T21	M51	2.34	5	**0.033**
T241	M51	2.64	11	**0.011**
T241	M55	2.88	10	**0.008**
T315	M35	10.19	11	**<.001**
T33	M55	14.66	11	**<.001**
T45	M35	15.1	11	**<.001**
T21	NV214	0.49	11	0.317
T21	NV22	11.73	11	**<.001**
T241	NV44	20.26	11	**<.001**
T315	NV214	1.06	10	0.157
T315	NV35	4.85	11	**<.001**
T315	NV415	2.61	9	**0.014**
T33	NV22	11.54	11	**<.001**
T33	NV44	2.63	11	**0.012**
T45	NV35	4.81	11	**<.001**
T45	NV415	3.34	7	**0.006**
T21	SV22	8.06	11	**<.001**
T241	SV44	11.38	11	**<.001**
T315	SV214	7.78	11	**<.001**
T33	SV22	8.87	11	**<.001**
T33	SV44	4.04	10	**0.001**
T45	SV21	1.35	9	0.105
T45	SV214	2.25	8	**0.027**
T45	SV35	1.8	10	**0.05**

Table A.4 Mixed effect models of native category choices for each Thai tone stimulus (significant results are in bold, $p < .05$)

Native language	Thai stimulus	*F*	*df*		*p*
Mandarin	T315	98.73	3	33	**<.001**
Mandarin	T241	16.78	3	33	**<.001**
Mandarin	T21	9	3	33	**<.001**
Mandarin	T33	259.64	3	33	**<.001**
Mandarin	T45	209.54	3	33	**<.001**
Northern Vietnamese	T315	22.92	5	55	**<.001**
Northern Vietnamese	T241	391.35	5	55	**<.001**
Northern Vietnamese	T21	106.4	5	55	**<.001**
Northern Vietnamese	T33	107.08	5	55	**<.001**
Northern Vietnamese	T45	16.33	5	55	**<.001**
Southern Vietnamese	T315	47.97	4	44	**<.001**
Southern Vietnamese	T241	140.69	4	44	**<.001**
Southern Vietnamese	T21	60.5	4	44	**<.001**
Southern Vietnamese	T33	68.65	4	44	**<.001**
Southern Vietnamese	T45	4.21	4	44	**0.006**

Table A.5 Multiple comparisons between native category choices for each Thai stimulus type with Tukey adjustment (significant results are in bold, $p < .05$). Only the most relevant comparisons are listed here.

Thai stimulus	Native choice pairs	*t*	*df*	*p*
T21	M214 - M51	2.62	33	0.060
T315	M35 - M214	11.05	33	**<.001**
T45	M35 - M214	18.65	33	**<.001**
T33	M55 - M51	22.23	33	**<.001**
T241	M55 - M51	-0.27	33	0.99
T21	NV214 - NV22	-13.89	55	**<.001**
T315	NV35 - NV415	4.12	55	**0.002**
T45	NV35 - NV415	4.58	55	**<.001**
T33	NV44 - NV22	-10.32	55	**<.001**
T241	NV44 - NV22	33.2	55	**<.001**
T21	SV22 - SV21	12.04	44	**<.001**
T315	SV35 - SV214	-10.16	44	**<.001**
T45	SV35 - SV214	-0.01	44	1
T45	SV214 - SV21	0.02	44	1
T45	SV35 - SV21	0.005	44	1
T33	SV44 - SV22	-5.48	44	**<.001**
T241	SV44 - SV21	18.75	44	**<.001**

Appendix B Supplementary material for Chapter 6

Statistical tests for determining perceptual assimilation types for Mandarin listeners.

Table B.1 *T*-tests of response categories against chance level 25%. Significant findings (p <.05) are shown in bold.

Memory load	Thai stimuli	Responses	*t*	*df*	*p*
Low	T21	M214	1.60	10	.070
Low	T21	M51	4.34	14	**< .001**
Low	T33	M55	16.00	13	**< .001**
Low	T45	M35	12.49	15	**< .001**
Low	T241	M51	5.92	15	**< .001**
Low	T315	M35	3.61	14	**.001**
Low	T315	M214	4.04	14	**< .001**
High	T21	M55	1.34	10	.104
High	T21	M51	6.30	14	< .001
High	T33	M55	14.29	14	< .001
High	T45	M35	9.80	15	< .001
High	T241	M55	2.08	11	.031
High	T241	M51	7.67	15	< .001
High	T315	M35	3.46	14	.001
High	T315	M214	4.63	15	< .001

Table B.2 Linear mixed-effect models on native response choices for Thai tones by Mandarin listeners, conducted to determine whether native response categories were selected with different frequency for each Thai tone.

Memory load	Thai stimuli	*F*	*df*		*p*
Low	T21	11.61	3	45	< .001
High	T21	24.57	3	45	< .001
Low	T33	35.92	3	45	< .001
High	T33	67.37	3	45	< .001
Low	T45	145.67	3	45	< .001
High	T45	87.87	3	45	< .001
Low	T241	35.63	3	45	< .001
High	T241	60.51	3	45	< .001
Low	T315	28.16	3	45	< .001
High	T315	38.42	3	45	< .001

Table B.3 Pairwise comparisons (with Tukey adjustments) among native response choices for Thai tones by Mandarin listeners. Significant findings (p <.05) are shown in bold.

Memory load	Thai tones	Mandarin choice contrasts	*t*	*df*	*p*
Low	T21	M55 - M35	2.091	45	.172
Low	T21	M55 - M214	-0.664	45	.910
Low	T21	M55 - M51	-3.733	45	**.003**
Low	T21	M35 - M214	-2.756	45	**.041**
Low	T21	M35 - M51	-5.824	45	**< .001**
Low	T21	M214 - M51	-3.068	45	**.018**
High	T21	M55 - M35	2.995	45	**.022**
High	T21	M55 - M214	2.383	45	.095
High	T21	M55 - M51	-4.723	45	**< .001**
High	T21	M35 - M214	-0.613	45	.928
High	T21	M35 - M51	-7.718	45	**< .001**
High	T21	M214 - M51	-7.106	45	**< .001**
Low	T33	M55 - M35	8.813	45	**< .001**
Low	T33	M55 - M214	9.059	45	**< .001**
Low	T33	M55 - M51	6.863	45	**< .001**
Low	T33	M35 - M214	0.246	45	.995
Low	T33	M35 - M51	-1.950	45	.222
Low	T33	M214 - M51	-2.196	45	.140
High	T33	M55 - M35	12.184	45	**< .001**
High	T33	M55 - M214	11.976	45	**< .001**
High	T33	M55 - M51	10.289	45	**< .001**
High	T33	M35 - M214	-0.208	45	.997
High	T33	M35 - M51	-1.896	45	.244
High	T33	M214 - M51	-1.688	45	.342
Low	T45	M55 - M35	-17.622	45	**< .001**
Low	T45	M55 - M214	-1.960	45	.218
Low	T45	M55 - M51	0.000	45	1.000
Low	T45	M35 - M214	15.662	45	**< .001**
Low	T45	M35 - M51	17.622	45	**< .001**
Low	T45	M214 - M51	1.960	45	.218
High	T45	M55 - M35	-13.878	45	**< .001**
High	T45	M55 - M214	-2.404	45	.091
High	T45	M55 - M51	-0.036	45	1.000
High	T45	M35 - M214	11.474	45	**< .001**
High	T45	M35 - M51	13.842	45	**< .001**
High	T45	M214 - M51	2.368	45	.098
Low	T241	M55 - M35	2.270	45	.120
Low	T241	M55 - M214	2.569	45	.063

Low	T241	M55 - M51	-6.510	45	**< .001**
Low	T241	M35 - M214	0.299	45	.991
Low	T241	M35 - M51	-8.781	45	**< .001**
Low	T241	M214 - M51	-9.080	45	**< .001**
High	T241	M55 - M35	4.586	45	**< .001**
High	T241	M55 - M214	4.547	45	**< .001**
High	T241	M55 - M51	-7.079	45	**< .001**
High	T241	M35 - M214	-0.039	45	1.000
High	T241	M35 - M51	-11.665	45	**< .001**
High	T241	M214 - M51	-11.626	45	**< .001**
Low	T315	M55 - M35	-6.340	45	**< .001**
Low	T315	M55 - M214	-6.678	45	**< .001**
Low	T315	M55 - M51	-0.029	45	1.000
Low	T315	M35 - M214	-0.338	45	.987
Low	T315	M35 - M51	6.311	45	**< .001**
Low	T315	M214 - M51	6.649	45	**< .001**
High	T315	M55 - M35	-6.621	45	**< .001**
High	T315	M55 - M214	-8.335	45	**< .001**
High	T315	M55 - M51	0.034	45	1.000
High	T315	M35 - M214	-1.714	45	.329
High	T315	M35 - M51	6.654	45	**< .001**
High	T315	M214 - M51	8.368	45	**< .001**

Vietnamese listeners

Table B.4 *T*-tests of response categories against chance level 20%. Significant findings (p <.05) are shown in bold.

Memory load	Thai stimuli	Responses	*t*	*df*	*p*
Low	T21	SV22	16.875	15	**< .001**
Low	T33	SV44	2.877	14	**.006**
Low	T33	SV22	3.713	14	**.001**
Low	T45	SV35	2.583	8	**.016**
Low	T45	SV214	1.883	11	**.043**
Low	T45	SV21	2.831	13	**.007**
Low	T241	SV44	14.036	15	**< .001**
Low	T315	SV214	12.688	15	**< .001**
High	T21	SV22	33.765	15	**< .001**
High	T33	SV44	3.757	10	**.002**
High	T33	SV22	4.505	15	**< .001**
High	T45	SV214	1.239	11	.120
High	T45	SV21	5.783	13	**< .001**
High	T241	SV44	10.435	15	**< .001**
High	T315	SV214	26.457	15	**< .001**

Table B.5 Linear mixed-effect models on native response choices of Thai tones by Vietnamese listeners. (This is to determine whether native response categories were selected with different frequency for each Thai tone.)

Memory load	Thai stimuli	*F*	*df*		*p*
Low	T21	265.76	4	60	< .001
High	T21	1061.70	4	60	< .001
Low	T33	18.22	4	60	< .001
High	T33	23.87	4	60	< .001
Low	T45	6.68	4	60	< .001
High	T45	18.53	4	60	< .001
Low	T241	170.63	4	60	< .001
High	T241	85.80	4	60	< .001
Low	T315	128.42	4	60	< .001
High	T315	549.57	4	60	< .001

Table B.6 Pairwise comparisons (with Tukey adjustments) between native response choices of Thai tones by Vietnamese listeners. Significant findings (p <.05) are shown in bold.

Memory load	Thai stimuli	SV contrasts	*t*	*df*	*p*
Low	T21	SV44 - SV35	1.1484	60	.780
Low	T21	SV44 - SV22	-25.2750	60	**< .001**
Low	T21	SV44 - SV214	1.2202	60	.740
Low	T21	SV44 - SV21	-0.5965	60	.975
Low	T21	SV35 - SV22	-26.4234	60	**< .001**
Low	T21	SV35 - SV214	0.0718	60	1.000
Low	T21	SV35 - SV21	-1.7449	60	.415
Low	T21	SV22 - SV214	26.4952	60	**< .001**
Low	T21	SV22 - SV21	24.6785	60	**< .001**
Low	T21	SV214 - SV21	-1.8167	60	.374
High	T21	SV44 - SV35	1.8759	60	.341
High	T21	SV44 - SV22	-50.5006	60	**< .001**
High	T21	SV44 - SV214	1.1227	60	.794
High	T21	SV44 - SV21	0.9934	60	.857
High	T21	SV35 - SV22	-52.3764	60	**< .001**
High	T21	SV35 - SV214	-0.7531	60	.943
High	T21	SV35 - SV21	-0.8825	60	.902
High	T21	SV22 - SV214	51.6233	60	**< .001**
High	T21	SV22 - SV21	51.4939	60	**< .001**
High	T21	SV214 - SV21	-0.1294	60	0.999
Low	T33	SV44 - SV35	5.0324	60	**< .001**
Low	T33	SV44 - SV22	-0.9763	60	.865
Low	T33	SV44 - SV214	5.1659	60	**< .001**
Low	T33	SV44 - SV21	4.7391	60	**< .001**
Low	T33	SV35 - SV22	-6.0087	60	**< .001**
Low	T33	SV35 - SV214	0.1334	60	.999
Low	T33	SV35 - SV21	-0.2934	60	.998
Low	T33	SV22 - SV214	6.1421	60	**< .001**
Low	T33	SV22 - SV21	5.7153	60	**< .001**
Low	T33	SV214 - SV21	-0.4268	60	.993
High	T33	SV44 - SV35	4.6190	60	**< .001**
High	T33	SV44 - SV22	-2.8467	60	**.046**
High	T33	SV44 - SV214	4.6190	60	**< .001**
High	T33	SV44 - SV21	4.5920	60	**< .001**
High	T33	SV35 - SV22	-7.4658	60	**< .001**
High	T33	SV35 - SV214	<.001	60	1.000
High	T33	SV35 - SV21	-0.0270	60	1.000
High	T33	SV22 - SV214	7.4658	60	**< .001**

High	T33	SV22 - SV21	7.4387	60	**< .001**
High	T33	SV214 - SV21	-0.0270	60	1.000
Low	T45	SV44 - SV35	-2.0542	60	.254
Low	T45	SV44 - SV22	0.4589	60	.991
Low	T45	SV44 - SV214	-2.0006	60	.278
Low	T45	SV44 - SV21	-4.0920	60	**.001**
Low	T45	SV35 - SV22	2.5131	60	.101
Low	T45	SV35 - SV214	0.0537	60	1.000
Low	T45	SV35 - SV21	-2.0377	60	.261
Low	T45	SV22 - SV214	-2.4595	60	.114
Low	T45	SV22 - SV21	-4.5509	60	**< .001**
Low	T45	SV214 - SV21	-2.0914	60	.237
High	T45	SV44 - SV35	-1.8336	60	.364
High	T45	SV44 - SV22	<.001	60	1.000
High	T45	SV44 - SV214	-2.7353	60	**.060**
High	T45	SV44 - SV21	-7.4116	60	**< .001**
High	T45	SV35 - SV22	1.8336	60	.364
High	T45	SV35 - SV214	-0.9016	60	.895
High	T45	SV35 - SV21	-5.5779	60	**< .001**
High	T45	SV22 - SV214	-2.7353	60	**.060**
High	T45	SV22 - SV21	-7.4116	60	**< .001**
High	T45	SV214 - SV21	-4.6763	60	**< .001**
Low	T241	SV44 - SV35	21.0853	60	**< .001**
Low	T241	SV44 - SV22	18.1082	60	**< .001**
Low	T241	SV44 - SV214	21.3871	60	**< .001**
Low	T241	SV44 - SV21	21.1521	60	**< .001**
Low	T241	SV35 - SV22	-2.9771	60	**.033**
Low	T241	SV35 - SV214	0.3018	60	.998
Low	T241	SV35 - SV21	0.0668	60	1.000
Low	T241	SV22 - SV214	3.2789	60	**.014**
Low	T241	SV22 - SV21	3.0439	60	**.028**
Low	T241	SV214 - SV21	-0.2350	60	.999
High	T241	SV44 - SV35	14.9045	60	**< .001**
High	T241	SV44 - SV22	12.1165	60	**< .001**
High	T241	SV44 - SV214	15.2415	60	**< .001**
High	T241	SV44 - SV21	15.1576	60	**< .001**
High	T241	SV35 - SV22	-2.7880	60	**.053**
High	T241	SV35 - SV214	0.3370	60	.997
High	T241	SV35 - SV21	0.2532	60	.999
High	T241	SV22 - SV214	3.1250	60	**.022**
High	T241	SV22 - SV21	3.0412	60	**.028**
High	T241	SV214 - SV21	-0.0839	60	1.000
Low	T315	SV44 - SV35	-1.6296	60	.485

Low	TT315	SV44 - SV22	-0.3445	60	.997
Low	T315	SV44 - SV214	-18.6478	60	**< .001**
Low	T315	SV44 - SV21	-1.1812	60	.762
Low	T315	SV35 - SV22	1.2851	60	.701
Low	T315	SV35 - SV214	-17.0182	60	**< .001**
Low	T315	SV35 - SV21	0.4484	60	.991
Low	T315	SV22 - SV214	-18.3033	60	**< .001**
Low	T315	SV22 - SV21	-0.8367	60	.918
Low	T315	SV214 - SV21	17.4666	60	**< .001**
High	T315	SV44 - SV35	-0.5516	60	.981
High	T315	SV44 - SV22	0.0959	60	1.000
High	T315	SV44 - SV214	-37.7749	60	**< .001**
High	T315	SV44 - SV21	-2.7305	60	**.061**
High	T315	SV35 - SV22	0.6475	60	.967
High	T315	SV35 - SV214	-37.2233	60	**< .001**
High	T315	SV35 - SV21	-2.1789	60	.202
High	T315	SV22 - SV214	-37.8708	60	**< .001**
High	T315	SV22 - SV21	-2.8264	60	**.048**
High	T315	SV214 - SV21	35.0444	60	**< .001**

Mean and confidence intervals of the overlap scores and fit index difference scores.

Table B.7 Mean and confidence intervals of the overlap scores and fit index difference scores of Thai tone contrasts as perceived by Mandarin listeners.

Memory load	Thai contrasts	Overlap scores			Fit index difference scores		
		Mean	95% CI		Mean	95%CI	
			LL	UL		LL	UL
Low	T33-T45	0.039	-0.006	0.083	1.750	1.640	1.860
	T33-T21	0.356	0.193	0.519	1.070	0.790	1.340
	T33-T241	0.433	0.248	0.619	1.060	0.722	1.410
	T241-T21	0.584	0.451	0.718	0.738	0.552	0.924
	T315-T45	0.589	0.464	0.714	0.782	0.534	1.030
High	T33-T45	0.009	-0.010	0.029	1.760	1.650	1.870
	T33-T21	0.409	0.231	0.587	0.988	0.732	1.240
	T33-T241	0.420	0.291	0.549	1.030	0.822	1.230
	T315-T45	0.589	0.480	0.699	0.802	0.601	1.000
	T241-T21	0.735	0.624	0.846	0.507	0.349	0.665

Note: for each participant per Thai tone, we calculate the sum of the overlap score and fit index difference score of all the response categories.

Table B.8 Mean and confidence intervals of the overlap scores and fit index difference scores of Thai tone contrasts as assimilated by Vietnamese listeners.

Memory load	Thai contrasts	Overlap scores			Fit index difference scores		
		Mean	95% CI		Mean	95%CI	
			LL	UL		LL	UL
Low	T33-T45	0.119	0.010	0.228	1.300	1.080	1.520
	T241-T21	0.189	0.094	0.284	1.270	1.100	1.440
	T315-T45	0.387	0.238	0.535	0.877	0.630	1.120
	T33-T21	0.584	0.401	0.768	0.642	0.367	0.916
	T33-T241	0.588	0.453	0.722	0.659	0.427	0.891
High	T33-T45	0.007	-0.004	0.017	1.250	1.100	1.390
	T241-T21	0.204	0.090	0.318	1.120	0.924	1.310
	T315-T45	0.312	0.160	0.463	0.839	0.630	1.050
	T33-T241	0.541	0.398	0.684	0.689	0.476	0.901
	T33-T21	0.650	0.475	0.825	0.510	0.281	0.739

Note: for each participant per Thai tone, we calculate the sum of the overlap score and fit index difference score of all the response categories.

Statistical tests for discrimination tasks

Table B.9 LMER model results for Mandarin listener discrimination.

Effects	Estimate	*SE*	*t*
Intercept	3.89	0.26	14.84
MemoryLoadLow	-0.06	0.37	-0.17
TalkerVariable	-0.91	0.33	-2.78
VowelVariable	-1.14	0.33	-3.49
T315-T45	-0.54	0.33	-1.65
T33-T21	0.13	0.33	0.40
T33-T241	-1.22	0.33	-3.74
T33-T45	0.96	0.33	2.94
MemoryLoadLow:TalkerVariable	0.20	0.46	0.43
MemoryLoadLow:VowelVariable	0.18	0.46	0.38
TalkerVariable:VowelVariable	0.78	0.46	1.69
MemoryLoadLow: T315-T45	-0.36	0.46	-0.79
MemoryLoadLow: T33-T21	0.04	0.46	0.08
MemoryLoadLow: T33-T241	0.01	0.46	0.02
MemoryLoadLow: T33-T45	0.16	0.46	0.35
TalkerVariable: T315-T45	-0.24	0.46	-0.52
TalkerVariable: T33-T21	0.27	0.46	0.57
TalkerVariable: T33-T241	0.19	0.46	0.42
TalkerVariable: T33-T45	0.19	0.46	0.40
VowelVariable: T315-T45	-0.15	0.46	-0.33
VowelVariable: T33-T21	0.28	0.46	0.60
VowelVariable: T33-T241	0.22	0.46	0.47
VowelVariable: T33-T45	0.60	0.46	1.31
MemoryLoadLow:TalkerVariable:VowelVariable	-0.55	0.65	-0.85
MemoryLoadLow:TalkerVariable: T315-T45	-0.14	0.65	-0.21
MemoryLoadLow:TalkerVariable: T33-T21	-0.10	0.65	-0.15
MemoryLoadLow:TalkerVariable: T33-T241	-0.29	0.65	-0.44

MemoryLoadLow:TalkerVariable: T33-T45	-0.16	0.65	-0.24
MemoryLoadLow:VowelVariable: T315-T45	0.37	0.65	0.57
MemoryLoadLow:VowelVariable: T33-T21	-0.40	0.65	-0.61
MemoryLoadLow:VowelVariable: T33-T241	0.06	0.65	0.10
MemoryLoadLow:VowelVariable: T33-T45	-0.35	0.65	-0.53
TalkerVariable:VowelVariable: T315-T45	0.07	0.65	0.11
TalkerVariable:VowelVariable: T33-T21	0.10	0.65	0.16
TalkerVariable:VowelVariable: T33-T241	0.06	0.65	0.10
TalkerVariable:VowelVariable: T33-T45	-0.40	0.65	-0.61
MemoryLoadLow:TalkerVariable:VowelVariable: T315-T45	0.38	0.92	0.41
MemoryLoadLow:TalkerVariable:VowelVariable: T33-T21	0.37	0.92	0.40
MemoryLoadLow:TalkerVariable:VowelVariable: T33-T241	0.49	0.92	0.53
MemoryLoadLow:TalkerVariable:VowelVariable: T33-T45	0.60	0.92	0.65

Table B.10 Multiple comparisons (with Tukey adjustments) of discrimination of the five Thai tone contrasts by Mandarin listeners. Assimilation types: Two-Category No overlap: T33-T45 and T33-T21; Category-Goodness No overlap: T241-T21; UnCategorised-Categorised Partial overlap: T315-T45; Two-Category No overlap/Partial overlap: T33-241. Significant findings (p <.05) are shown in bold.

Thai tone contrasts	estimate	*SE*	*df*	*t*	*p*
(T33-T45) - (T33-T21)	0.92	0.12	510	7.95	**< .01**
(T33-T45) - (T241-T21)	1.28	0.12	510	11.12	**< .01**
(T33-T45) - (T315-T45)	2.08	0.12	510	18.00	**< .01**
(T33-T45) - (T33-T241)	2.27	0.12	510	19.71	**< .01**
(T33-T21) - (T241-T21)	0.37	0.12	510	3.17	**0.01**
(T33-T21) - (T315-T45)	1.16	0.12	510	10.05	**< .01**
(T33-T21) - (T33-T241)	1.36	0.12	510	11.76	**< .01**
(T241-T21) - (T315-T45)	0.79	0.12	510	6.88	**< .01**
(T241-T21) - (T33-T241)	0.99	0.12	510	8.59	**< .01**
(T315-T45) - (T33-T241)	0.20	0.12	510	1.71	0.43

Table B.11 LMER model results for Vietnamese listener discrimination.

Effects	Estimate	*SE*	*t*
Intercept	4.54	0.27	17.11
MemoryLoadLow	0.30	0.38	0.80
TalkerVariable	-0.56	0.35	-1.61
VowelVariable	-1.22	0.35	-3.49
T315-T45	-0.63	0.35	-1.80
T33-T21	-0.87	0.35	-2.50
T33-T241	-1.34	0.35	-3.85
T33-T45	-2.02	0.35	-5.82
MemoryLoadLow:TalkerVariable	-0.03	0.49	-0.07
MemoryLoadLow:VowelVariable	0.52	0.49	1.06
TalkerVariable:VowelVariable	0.51	0.49	1.03
MemoryLoadLow: T315-T45	-0.27	0.49	-0.56
MemoryLoadLow: T33-T21	-0.54	0.49	-1.09
MemoryLoadLow: T33-T241	-0.49	0.49	-1.00
MemoryLoadLow: T33-T45	-0.01	0.49	-0.03
TalkerVariable: T315-T45	0.12	0.49	0.25
TalkerVariable: T33-T21	-0.41	0.49	-0.84
TalkerVariable: T33-T241	-0.52	0.49	-1.06
TalkerVariable: T33-T45	-0.32	0.49	-0.66
VowelVariable: T315-T45	0.56	0.49	1.14
VowelVariable: T33-T21	-0.06	0.49	-0.12
VowelVariable: T33-T241	-0.11	0.49	-0.22
VowelVariable: T33-T45	0.08	0.49	0.17
MemoryLoadLow:TalkerVariable:VowelVariable	0.19	0.70	0.28
MemoryLoadLow:TalkerVariable: T315-T45	0.08	0.70	0.12
MemoryLoadLow:TalkerVariable: T33-T21	0.13	0.70	0.19
MemoryLoadLow:TalkerVariable: T33-T241	0.37	0.70	0.54
MemoryLoadLow:TalkerVariable: T33-T45	-0.31	0.70	-0.44

MemoryLoadLow:VowelVariable: T315-T45	0.14	0.70	0.21
MemoryLoadLow:VowelVariable: T33-T21	-0.18	0.70	-0.27
MemoryLoadLow:VowelVariable: T33-T241	-0.25	0.70	-0.36
MemoryLoadLow:VowelVariable: T33-T45	0.05	0.70	0.08
TalkerVariable:VowelVariable: T315-T45	-0.02	0.70	-0.04
TalkerVariable:VowelVariable: T33-T21	0.06	0.70	0.08
TalkerVariable:VowelVariable: T33-T241	0.66	0.70	0.95
TalkerVariable:VowelVariable: T33-T45	0.57	0.70	0.81
MemoryLoadLow:TalkerVariable:VowelVariable: T315-T45	-0.55	0.98	-0.56
MemoryLoadLow:TalkerVariable:VowelVariable: T33-T21	-0.12	0.98	-0.12
MemoryLoadLow:TalkerVariable:VowelVariable: T33-T241	-0.78	0.98	-0.80
MemoryLoadLow:TalkerVariable:VowelVariable: T33-T45	-0.19	0.98	-0.19

Table B.12 Multiple comparisons (with Tukey adjustments) of discrimination of different Thai tone contrasts by Vietnamese listeners.

Assimilation types: T241-T21 (Two-Category No overlap, T33-T21 (UnCategorised-Categorised Partial overlap/Single-Category No overlap), T315-T45 (UnCategorised-Categorised Partial overlap/Two-Category No overlap), T33-T241(UnCategorised-Categorised Partial overlap/Two-Category Partial overlap), T33-T45 UnCategorised-UnCategorised No overlap/Two-Category No overlap. Significant findings (p <.05) are shown in bold.

Thai contrasts	Estimate	*SE*	*t*	*df*	*p*
(T33-T45) - (T241-T21)	0.44	0.12	510	3.57	**< .01**
(T33-T45) - (T315-T45)	1.39	0.12	510	11.27	**< .01**
(T33-T45) - (T33-T21)	1.80	0.12	510	14.65	**< .01**
(T33-T45) - (T33-T241)	2.10	0.12	510	17.05	**< .01**
(T241-T21) - (T315-T45)	0.95	0.12	510	7.70	**< .01**
(T241-T21) - (T33-T21)	1.36	0.12	510	11.08	**< .01**
(T241-T21) - (T33-T241)	1.66	0.12	510	13.48	**< .01**
(T315-T45) - (T33-T21)	0.42	0.12	510	3.38	**0.01**
(T315-T45) - (T33-T241)	0.71	0.12	510	5.78	**< .01**
(T33-T21) - (T33-T241)	0.30	0.12	510	2.40	0.12

Statistical tests for comparing discrimination of the five contrasts across Mandarin and Vietnamese listeners.

Table B.13 LMER model summary for discrimination across Mandarin and Vietnamese listeners.

Effects	Estimate	*SE*	*t*
(Intercept)	3.89	0.26	14.74
Vietnamese	0.06	0.37	0.16
MemoryLoadLow	-0.06	0.37	-0.17
VariableSpeaker	-0.91	0.34	-2.69
VariableVowel	-1.14	0.34	-3.37
T315-T45	-0.54	0.34	-1.60
T33-T21	0.13	0.34	0.38
T33-T241	-1.22	0.34	-3.62
T33-T45	0.96	0.34	2.84
Vietnamese:MemoryLoadLow	0.03	0.53	0.06
Vietnamese:VariableSpeaker	0.52	0.48	1.09
MemoryLoadLow:VariableSpeaker	0.20	0.48	0.42
Vietnamese:VariableVowel	1.15	0.48	2.41
MemoryLoadLow:VariableVowel	0.18	0.48	0.37
VariableSpeaker:VariableVowel	0.78	0.48	1.63
Vietnamese:T315-T45	0.04	0.48	0.07
Vietnamese:T33-T21	-1.06	0.48	-2.22
Vietnamese:T33-T241	0.08	0.48	0.18
Vietnamese:T33-T45	-0.06	0.48	-0.12
MemoryLoadLow:T315-T45	-0.36	0.48	-0.76
MemoryLoadLow:T33-T21	0.04	0.48	0.08
MemoryLoadLow:T33-T241	0.01	0.48	0.02
MemoryLoadLow:T33-T45	0.16	0.48	0.34
VariableSpeaker:T315-T45	-0.24	0.48	-0.50
VariableSpeaker:T33-T21	0.26	0.48	0.56
VariableSpeaker:T33-T241	0.19	0.48	0.41
VariableSpeaker:T33-T45	0.18	0.48	0.39
VariableVowel:T315-T45	-0.15	0.48	-0.32
VariableVowel:T33-T21	0.28	0.48	0.58
VariableVowel:T33-T241	0.22	0.48	0.46
VariableVowel:T33-T45	0.60	0.48	1.26
Vietnamese:MemoryLoadLow:VariableSpeaker	-0.25	0.67	-0.37
Vietnamese:MemoryLoadLow:VariableVowel	-0.84	0.68	-1.25
Vietnamese:VariableSpeaker:VariableVowel	-0.65	0.67	-0.97

MemoryLoadLow:VariableSpeaker:VariableVowel	-0.55	0.67	-0.82
Vietnamese:MemoryLoadLow:T315-T45	0.63	0.67	0.93
Vietnamese:MemoryLoadLow:T33-T21	0.18	0.67	0.27
Vietnamese:MemoryLoadLow:T33-T241	-0.27	0.67	-0.40
Vietnamese:MemoryLoadLow:T33-T45	-0.44	0.67	-0.65
Vietnamese:VariableSpeaker:T315-T45	-0.25	0.67	-0.37
Vietnamese:VariableSpeaker:T33-T21	-0.62	0.67	-0.92
Vietnamese:VariableSpeaker:T33-T241	-1.03	0.67	-1.53
Vietnamese:VariableSpeaker:T33-T45	-0.39	0.67	-0.58
MemoryLoadLow:VariableSpeaker:T315-T45	-0.14	0.67	-0.21
MemoryLoadLow:VariableSpeaker:T33-T21	-0.10	0.67	-0.15
MemoryLoadLow:VariableSpeaker:T33-T241	-0.29	0.67	-0.43
MemoryLoadLow:VariableSpeaker:T33-T45	-0.16	0.67	-0.23
Vietnamese:VariableVowel:T315-T45	-0.80	0.67	-1.18
Vietnamese:VariableVowel:T33-T21	-1.34	0.67	-1.99
Vietnamese:VariableVowel:T33-T241	-0.79	0.67	-1.17
Vietnamese:VariableVowel:T33-T45	-1.31	0.67	-1.94
MemoryLoadLow:VariableVowel:T315-T45	0.37	0.67	0.55
MemoryLoadLow:VariableVowel:T33-T21	-0.40	0.67	-0.59
MemoryLoadLow:VariableVowel:T33-T241	0.06	0.67	0.09
MemoryLoadLow:VariableVowel:T33-T45	-0.35	0.67	-0.51
VariableSpeaker:VariableVowel:T315-T45	0.07	0.67	0.11
VariableSpeaker:VariableVowel:T33-T21	0.10	0.67	0.15
VariableSpeaker:VariableVowel:T33-T241	0.06	0.67	0.09
VariableSpeaker:VariableVowel:T33-T45	-0.40	0.67	-0.59
Vietnamese:MemoryLoadLow:VariableSpeaker:VariableVowel	0.91	0.95	0.95
Vietnamese:MemoryLoadLow:VariableSpeaker:T315-T45	0.09	0.95	0.09
Vietnamese:MemoryLoadLow:VariableSpeaker:T33-T21	-0.19	0.95	-0.20
Vietnamese:MemoryLoadLow:VariableSpeaker:T33-T241	0.68	0.95	0.71
Vietnamese:MemoryLoadLow:VariableSpeaker:T33-T45	0.24	0.95	0.25
Vietnamese:MemoryLoadLow:VariableVowel:T315-T45	-0.04	0.95	-0.04
Vietnamese:MemoryLoadLow:VariableVowel:T33-T21	0.80	0.95	0.84
Vietnamese:MemoryLoadLow:VariableVowel:T33-T241	0.03	0.95	0.03
Vietnamese:MemoryLoadLow:VariableVowel:T33-T45	0.49	0.95	0.51
Vietnamese:VariableSpeaker:VariableVowel:T315-T45	0.44	0.95	0.46
Vietnamese:VariableSpeaker:VariableVowel:T33-T21	0.35	0.95	0.37
Vietnamese:VariableSpeaker:VariableVowel:T33-T241	0.89	0.95	0.93
Vietnamese:VariableSpeaker:VariableVowel:T33-T45	0.97	0.95	1.02
MemoryLoadLow:VariableSpeaker:VariableVowel:T315-T45	0.38	0.95	0.40
MemoryLoadLow:VariableSpeaker:VariableVowel:T33-T21	0.36	0.95	0.38

MemoryLoadLow:VariableSpeaker:VariableVowel:T33-T241	0.49	0.95	0.51
MemoryLoadLow:VariableSpeaker:VariableVowel:T33-T45	0.60	0.95	0.63
Vietnamese:MemoryLoadLow:VariableSpeaker:VariableVowel:T315-T45	-0.81	1.35	-0.60
Vietnamese:MemoryLoadLow:VariableSpeaker:VariableVowel:T33-T21	-0.13	1.35	-0.10
Vietnamese:MemoryLoadLow:VariableSpeaker:VariableVowel:T33-T241	-0.85	1.35	-0.63
Vietnamese:MemoryLoadLow:VariableSpeaker:VariableVowel:T33-T45	-1.15	1.35	-0.85

Table B.14 Effects for discrimination comparisons across Mandarin and Vietnamese listeners. Significant findings (p <.05) are shown in bold.

Effects	*F*	*df*		*p*
Language	0.54	1	60	0.46
Memory load	0.55	1	60	0.46
Talker	71.07	1	60	**<.01**
Vowel	101.33	1	60	**<.01**
ToneContrast	198.98	4	1020	**<.01**
Language:Memory load	0.10	1	60	0.75
Language:Talker	0.12	1	60	0.73
Memory load:Talker	0.01	1	60	0.94
Language:Vowel	0.07	1	60	0.79
Memory load:Vowel	3.06	1	60	**0.09**
Talker:Vowel	39.80	1	1020	**<.01**
Language: ToneContrast	32.80	4	1020	**<.01**
Memory load: ToneContrast	0.57	4	1020	0.68
Talker: ToneContrast	1.24	4	1020	0.29
Vowel: ToneContrast	1.52	4	1020	0.19
Language:Memory load:Talker	0.13	1	60	0.72
Language:Memory load:Vowel	3.81	1	60	**0.06**
Language:Talker:Vowel	0.03	1	1020	0.86
Memory load:Talker:Vowel	0.01	1	1020	0.90
Language:Memory load: ToneContrast	2.26	4	1020	**0.06**
Language:Talker: ToneContrast	1.09	4	1020	0.36
Memory load:Talker: ToneContrast	0.08	4	1020	0.99
Language:Vowel: ToneContrast	2.26	4	1020	**0.06**
Memory load:Vowel: ToneContrast	0.46	4	1020	0.76
Talker:Vowel: ToneContrast	0.88	4	1020	0.48
Language:Memory load:Talker:Vowel	0.56	1	1020	0.45
Language:Memory load:Talker: ToneContrast	0.28	4	1020	0.89
Language:Memory load:Vowel: ToneContrast	0.98	4	1020	0.42
Language:Talker:Vowel: ToneContrast	0.20	4	1020	0.94
Memory load:Talker:Vowel: ToneContrast	0.08	4	1020	0.99
Language:Memory load:Talker:Vowel: ToneContrast	0.27	4	1020	0.90

Table B.15 Cross-language comparisons of five Thai tone contrasts (with Tukey adjustments). Significant findings (p <.05) are shown in bold.

Contrast				Estimate	*SE*	*df*	*t*	*p*
Mandarin	T33-T45	Vietnamese	T33-T45	0.26	0.17	153	1.48	0.90
Mandarin	T33-T45	Vietnamese	T33-T21	2.06	0.17	153	11.88	**<.01**
Mandarin	T33-T45	Vietnamese	T241-T21	0.69	0.17	153	4.01	**<.01**
Mandarin	T33-T45	Vietnamese	T315-T45	1.64	0.17	153	9.48	**<.01**
Mandarin	T33-T45	Vietnamese	T33-T241	2.35	0.17	153	13.59	**<.01**
Mandarin	T33-T21	Vietnamese	T33-T21	1.14	0.17	153	6.58	**<.01**
Mandarin	T33-T21	Vietnamese	T241-T21	-0.22	0.17	153	-1.29	0.96
Mandarin	T33-T21	Vietnamese	T315-T45	0.72	0.17	153	4.18	**<.01**
Mandarin	T33-T21	Vietnamese	T33-T241	1.44	0.17	153	8.29	**<.01**
Mandarin	T241-T21	Vietnamese	T241-T21	-0.59	0.17	153	-3.40	**0.03**
Mandarin	T241-T21	Vietnamese	T315-T45	0.36	0.17	153	2.07	0.55
Mandarin	T241-T21	Vietnamese	T33-T241	1.07	0.17	153	6.18	**<.01**
Mandarin	T315-T45	Vietnamese	T315-T45	-0.44	0.17	153	-2.52	0.27
Mandarin	T315-T45	Vietnamese	T33-T241	0.28	0.17	153	1.59	0.85
Mandarin	T33-T241	Vietnamese	T33-T241	0.08	0.17	153	0.45	1.00

Appendix C Supplementary materials for Chapter 7

Acoustic measures of lexical tones

Table C.1 Acoustic measures of tones in Thai (20 tokens per tone), Mandarin and Vietnamese (32 tokens per tone). $F0_{mean}$, $F0_{excursion,}$ are Lobanov-normalised Hz scores (Lobanov, 1971).

	Duration (ms)			$F0_{mean}$			$F0_{excursion}$			$F0_{maxloc}$		
Tones	Mean	95% CI		Mean	95% CI		Mean	95% CI		Mean	95% CI	
		LL	UL		LL	UL		LL	UL		LL	UL
T45	619	597	641	0.05	0.03	0.07	0.17	0.15	0.19	1.00	0.99	1.00
T33	640	613	667	-0.01	-0.04	0.02	0.11	0.09	0.13	0.34	0.27	0.42
T21	622	598	646	-0.10	-0.11	-0.08	0.22	0.19	0.25	0.22	0.19	0.24
T315	642	616	669	-0.08	-0.10	-0.07	0.19	0.16	0.22	0.72	0.59	0.85
T241	565	533	597	0.13	0.11	0.15	0.21	0.18	0.24	0.53	0.49	0.58
M55	663	628	697	0.14	0.13	0.15	0.09	0.08	0.10	0.74	0.67	0.81
M35	613	587	640	-0.05	-0.07	-0.04	0.30	0.29	0.32	0.99	0.99	1.00
M214	745	711	779	-0.23	-0.26	-0.21	0.43	0.37	0.50	0.42	0.34	0.50
M51	506	474	538	0.08	0.07	0.10	0.44	0.39	0.49	0.29	0.26	0.32
V44	469	460	479	0.08	0.07	0.09	0.16	0.14	0.19	0.58	0.50	0.65
V22	501	491	511	-0.11	-0.12	-0.10	0.17	0.16	0.19	0.16	0.14	0.18
V35	465	455	475	0.19	0.18	0.21	0.44	0.42	0.47	1.00	0.99	1.00
V21	456	446	466	-0.17	-0.18	-0.15	0.26	0.20	0.32	0.49	0.38	0.60
V214	488	479	497	-0.09	-0.10	-0.07	0.48	0.46	0.51	1.00	1.00	1.00

Multiple comparisons of fixed factors for Mandarin data.

Table C.2 LMER model results for Mandarin listener imitation in terms of duration.

Effects	Estimate	*SE*	*t*
(Intercept)	-0.0200	0.01	-2.001
LowMemoryLoad	-0.0200	0.01	-1.115
ConstantTalker	0.0500	0.01	3.466
ConstantVowel	0.0100	0.01	0.681
Tone33	0.1100	0.01	7.623
Tone45	0.0500	0.01	3.617
Tone241	0.0800	0.01	5.936
Tone315	-0.0200	0.01	-1.543
LowMemoryLoad:ConstantTalker	0.0000	0.02	-0.144
LowMemoryLoad:ConstantVowel	0.0000	0.02	-0.149
ConstantTalker:ConstantVowel	-0.0100	0.02	-0.309
LowMemoryLoad:Tone33	-0.0200	0.02	-0.951
LowMemoryLoad:Tone45	0.0100	0.02	0.670
LowMemoryLoad:Tone241	0.0000	0.02	-0.087
LowMemoryLoad:Tone315	0.0300	0.02	1.486
ConstantTalker:Tone33	-0.0800	0.02	-4.163
ConstantTalker:Tone45	-0.0700	0.02	-3.592
ConstantTalker:Tone241	-0.0500	0.02	-2.356
ConstantTalker:Tone315	-0.0600	0.02	-2.881
ConstantVowel:Tone33	-0.0400	0.02	-2.050
ConstantVowel:Tone45	-0.0300	0.02	-1.390
ConstantVowel:Tone241	-0.0300	0.02	-1.278
ConstantVowel:Tone315	-0.0200	0.02	-0.956
LowMemoryLoad:ConstantTalker:ConstantVowel	-0.0100	0.03	-0.381
LowMemoryLoad:ConstantTalker:Tone33	0.0500	0.03	1.768
LowMemoryLoad:ConstantTalker:Tone45	0.0100	0.03	0.490
LowMemoryLoad:ConstantTalker:Tone241	0.0200	0.03	0.663

LowMemoryLoad:ConstantTalker:Tone315	0.0200	0.03	0.717
LowMemoryLoad:ConstantVowel:Tone33	0.0400	0.03	1.546
LowMemoryLoad:ConstantVowel:Tone45	0.0200	0.03	0.825
LowMemoryLoad:ConstantVowel:Tone241	0.0200	0.03	0.565
LowMemoryLoad:ConstantVowel:Tone315	0.0100	0.03	0.519
ConstantTalker:ConstantVowel:Tone33	0.0400	0.03	1.414
ConstantTalker:ConstantVowel:Tone45	0.0200	0.03	0.706
ConstantTalker:ConstantVowel:Tone241	0.0100	0.03	0.225
ConstantTalker:ConstantVowel:Tone315	0.0100	0.03	0.295
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone33	-0.0500	0.04	-1.252
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone45	-0.0200	0.04	-0.383
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone241	0.0000	0.04	-0.112
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone315	0.0100	0.04	0.251

Table C.3 LMER model results for Mandarin listener imitation in terms of $F0_{mean}$.

Effects	Estimate	SE	t
(Intercept)	-0.0007	0.01	-0.119
LowMemoryLoad	0.0154	0.01	1.842
ConstantTalker	-0.0034	0.01	-0.425
ConstantVowel	-0.0033	0.01	-0.419
Tone33	-0.0092	0.01	-1.176
Tone45	-0.0064	0.01	-0.820
Tone241	-0.0505	0.01	-6.414
Tone315	0.0603	0.01	7.675
LowMemoryLoad:ConstantTalker	-0.0082	0.01	-0.734
LowMemoryLoad:ConstantVowel	-0.0031	0.01	-0.278
ConstantTalker:ConstantVowel	-0.0066	0.01	-0.589
LowMemoryLoad:Tone33	-0.0128	0.01	-1.154
LowMemoryLoad:Tone45	-0.0395	0.01	-3.547
LowMemoryLoad:Tone241	-0.0115	0.01	-1.030
LowMemoryLoad:Tone315	-0.0407	0.01	-3.662
ConstantTalker:Tone33	0.0038	0.01	0.339
ConstantTalker:Tone45	-0.0024	0.01	-0.220
ConstantTalker:Tone241	0.0317	0.01	2.845
ConstantTalker:Tone315	0.0000	0.01	0.000
ConstantVowel:Tone33	0.0017	0.01	0.155
ConstantVowel:Tone45	0.0020	0.01	0.178
ConstantVowel:Tone241	0.0108	0.01	0.968
ConstantVowel:Tone315	-0.0016	0.01	-0.140
LowMemoryLoad:ConstantTalker:ConstantVowel	0.0029	0.02	0.186
LowMemoryLoad:ConstantTalker:Tone33	0.0031	0.02	0.199
LowMemoryLoad:ConstantTalker:Tone45	0.0100	0.02	0.635
LowMemoryLoad:ConstantTalker:Tone241	0.0022	0.02	0.142
LowMemoryLoad:ConstantTalker:Tone315	0.0131	0.02	0.833

LowMemoryLoad:ConstantVowel:Tone33	0.0033	0.02	0.211
LowMemoryLoad:ConstantVowel:Tone45	0.0038	0.02	0.240
LowMemoryLoad:ConstantVowel:Tone241	-0.0064	0.02	-0.405
LowMemoryLoad:ConstantVowel:Tone315	0.0103	0.02	0.653
ConstantTalker:ConstantVowel:Tone33	-0.0019	0.02	-0.119
ConstantTalker:ConstantVowel:Tone45	0.0108	0.02	0.687
ConstantTalker:ConstantVowel:Tone241	-0.0019	0.02	-0.121
ConstantTalker:ConstantVowel:Tone315	0.0099	0.02	0.626
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone33	0.0067	0.02	0.300
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone45	-0.0029	0.02	-0.128
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone241	0.0018	0.02	0.080
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone315	-0.0088	0.02	-0.396

Table C.4 LMER model results for Mandarin listener imitation in terms of $F0_{excursion}$.

Effects	Estimate	SE	t
(Intercept)	0.0550	0.02	2.701
LowMemoryLoad	-0.0163	0.03	-0.566
ConstantTalker	0.0079	0.02	0.438
ConstantVowel	-0.0005	0.02	-0.028
Tone33	0.0767	0.02	4.278
Tone45	0.0226	0.02	1.264
Tone241	0.0597	0.02	3.325
Tone315	0.1517	0.02	8.469
LowMemoryLoad:ConstantTalker	0.0132	0.03	0.518
LowMemoryLoad:ConstantVowel	-0.0148	0.03	-0.583
ConstantTalker:ConstantVowel	-0.0036	0.03	-0.140
LowMemoryLoad:Tone33	-0.0325	0.03	-1.280
LowMemoryLoad:Tone45	0.0078	0.03	0.308
LowMemoryLoad:Tone241	-0.0305	0.03	-1.201
LowMemoryLoad:Tone315	-0.0543	0.03	-2.142
ConstantTalker:Tone33	0.0224	0.03	0.881
ConstantTalker:Tone45	0.0200	0.03	0.788
ConstantTalker:Tone241	-0.0178	0.03	-0.700
ConstantTalker:Tone315	-0.0130	0.03	-0.511
ConstantVowel:Tone33	0.0205	0.03	0.809
ConstantVowel:Tone45	0.0283	0.03	1.117
ConstantVowel:Tone241	0.0189	0.03	0.746
ConstantVowel:Tone315	0.0057	0.03	0.226
LowMemoryLoad:ConstantTalker:ConstantVowel	0.0274	0.04	0.762
LowMemoryLoad:ConstantTalker:Tone33	-0.0263	0.04	-0.732
LowMemoryLoad:ConstantTalker:Tone45	-0.0455	0.04	-1.266
LowMemoryLoad:ConstantTalker:Tone241	0.0179	0.04	0.499
LowMemoryLoad:ConstantTalker:Tone315	-0.0115	0.04	-0.320

LowMemoryLoad:ConstantVowel:Tone33	-0.0088	0.04	-0.246
LowMemoryLoad:ConstantVowel:Tone45	-0.0216	0.04	-0.603
LowMemoryLoad:ConstantVowel:Tone241	0.0161	0.04	0.450
LowMemoryLoad:ConstantVowel:Tone315	0.0198	0.04	0.551
ConstantTalker:ConstantVowel:Tone33	-0.0450	0.04	-1.252
ConstantTalker:ConstantVowel:Tone45	-0.0387	0.04	-1.077
ConstantTalker:ConstantVowel:Tone241	-0.0011	0.04	-0.030
ConstantTalker:ConstantVowel:Tone315	0.0091	0.04	0.254
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone33	-0.0002	0.05	-0.005
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone45	0.0118	0.05	0.233
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone241	-0.0339	0.05	-0.667
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone315	-0.0575	0.05	-1.133

Table C.5 LMER model results for Mandarin listener imitation in terms of $F0_{maxloc}$.

Effects	Estimate	SE	t
(Intercept)	0.0078	0.02	0.363
LowMemoryLoad	0.0008	0.03	0.026
ConstantTalker	0.0532	0.03	2.081
ConstantVowel	-0.0049	0.03	-0.192
Tone33	0.0527	0.03	2.073
Tone45	-0.0977	0.03	-3.838
Tone241	-0.1490	0.03	-5.846
Tone315	0.1660	0.03	6.525
LowMemoryLoad:ConstantTalker	-0.0237	0.04	-0.656
LowMemoryLoad:ConstantVowel	-0.0076	0.04	-0.212
ConstantTalker:ConstantVowel	-0.0221	0.04	-0.613
LowMemoryLoad:Tone33	-0.0281	0.04	-0.781
LowMemoryLoad:Tone45	0.0314	0.04	0.871
LowMemoryLoad:Tone241	0.0037	0.04	0.103
LowMemoryLoad:Tone315	-0.0369	0.04	-1.025
ConstantTalker:Tone33	-0.1791	0.04	-4.963
ConstantTalker:Tone45	-0.0375	0.04	-1.041
ConstantTalker:Tone241	-0.0795	0.04	-2.200
ConstantTalker:Tone315	-0.2573	0.04	-7.135
ConstantVowel:Tone33	0.0098	0.04	0.272
ConstantVowel:Tone45	0.0234	0.04	0.651
ConstantVowel:Tone241	-0.0008	0.04	-0.021
ConstantVowel:Tone315	0.0025	0.04	0.070
LowMemoryLoad:ConstantTalker:ConstantVowel	0.0181	0.05	0.355
LowMemoryLoad:ConstantTalker:Tone33	0.0002	0.05	0.004
LowMemoryLoad:ConstantTalker:Tone45	0.0206	0.05	0.403
LowMemoryLoad:ConstantTalker:Tone241	-0.0209	0.05	-0.409
LowMemoryLoad:ConstantTalker:Tone315	0.0559	0.05	1.097

LowMemoryLoad:ConstantVowel:Tone33	-0.0128	0.05	-0.252
LowMemoryLoad:ConstantVowel:Tone45	-0.0148	0.05	-0.291
LowMemoryLoad:ConstantVowel:Tone241	-0.0131	0.05	-0.257
LowMemoryLoad:ConstantVowel:Tone315	0.0081	0.05	0.158
ConstantTalker:ConstantVowel:Tone33	0.0095	0.05	0.185
ConstantTalker:ConstantVowel:Tone45	-0.0076	0.05	-0.148
ConstantTalker:ConstantVowel:Tone241	-0.0007	0.05	-0.013
ConstantTalker:ConstantVowel:Tone315	0.0352	0.05	0.691
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone33	0.0560	0.07	0.776
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone45	0.0045	0.07	0.062
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone241	0.0580	0.07	0.803
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone315	-0.0253	0.07	-0.352

Table C.6 Multiple comparisons of tone main effect on imitations by Mandarin participants with Tukey adjustments. Effect sizes are shown, using Cohen's *d.* Significant findings (p <.05) are shown in bold.

Features	Contrasts	*d*	Estimates	*SE*	*df*	*t*	*p*
Duration	21 - 33	-0.664	-0.077	0.005	5026	-14.98	**<.001**
Duration	21 - 45	-0.178	-0.021	0.005	5026	-4.02	**0.001**
Duration	21 - 241	-0.616	-0.072	0.005	5026	-13.91	**<.001**
Duration	21 - 315	0.079	0.009	0.005	5026	1.79	0.377
Duration	33 - 45	0.485	0.056	0.005	5026	10.95	**<.001**
Duration	33 - 241	0.047	0.006	0.005	5026	1.07	0.822
Duration	33 - 315	0.743	0.086	0.005	5026	16.79	**<.001**
Duration	45 - 241	-0.438	-0.051	0.005	5026	-9.88	**<.001**
Duration	45 - 315	0.258	0.030	0.005	5026	5.82	**<.001**
Duration	241 - 315	0.696	0.081	0.005	5026	15.71	**<.001**
$F0_{mean}$	21 - 33	0.174	0.011	0.003	5026	3.92	**0.001**
$F0_{mean}$	21 - 45	0.328	0.021	0.003	5026	7.41	**<.001**
$F0_{mean}$	21 - 241	0.577	0.036	0.003	5026	13.02	**<.001**
$F0_{mean}$	21 - 315	-0.738	-0.046	0.003	5026	-16.66	**<.001**
$F0_{mean}$	33 - 45	0.155	0.010	0.003	5026	3.49	**0.004**
$F0_{mean}$	33 - 241	0.403	0.025	0.003	5026	9.1	**<.001**
$F0_{mean}$	33 - 315	-0.911	-0.057	0.003	5026	-20.58	**<.001**
$F0_{mean}$	45 - 241	0.249	0.016	0.003	5026	5.61	**<.001**
$F0_{mean}$	45 - 315	-1.066	-0.067	0.003	5026	-24.08	**<.001**
$F0_{mean}$	241 - 315	-1.315	-0.083	0.003	5026	-29.69	**<.001**
$F0_{excursion}$	21 - 33	-0.431	-0.062	0.006	5026	-9.73	**<.001**
$F0_{excursion}$	21 - 45	-0.179	-0.026	0.006	5026	-4.05	**0.001**
$F0_{excursion}$	21 - 241	-0.342	-0.049	0.006	5026	-7.72	**<.001**
$F0_{excursion}$	21 - 315	-0.824	-0.118	0.006	5026	-18.62	**<.001**
$F0_{excursion}$	33 - 45	0.252	0.036	0.006	5026	5.68	**<.001**
$F0_{excursion}$	33 - 241	0.089	0.013	0.006	5026	2.01	0.26

$F0_{excursion}$	33 - 315	-0.393	-0.056	0.006	5026	-8.87	**<.001**
$F0_{excursion}$	45 - 241	-0.163	-0.023	0.006	5026	-3.67	**0.002**
$F0_{excursion}$	45 - 315	-0.645	-0.092	0.006	5026	-14.56	**<.001**
$F0_{excursion}$	241 - 315	-0.482	-0.069	0.006	5026	-10.89	**<.001**
$F0_{maxloc}$	21 - 33	0.195	0.04	0.009	5026	4.41	**<.001**
$F0_{maxloc}$	21 - 45	0.437	0.089	0.009	5026	9.85	**<.001**
$F0_{maxloc}$	21 - 241	0.927	0.189	0.009	5026	20.92	**<.001**
$F0_{maxloc}$	21 - 315	-0.205	-0.042	0.009	5026	-4.64	**<.001**
$F0_{maxloc}$	33 - 45	0.241	0.049	0.009	5026	5.44	**<.001**
$F0_{maxloc}$	33 - 241	0.732	0.149	0.009	5026	16.51	**<.001**
$F0_{maxloc}$	33 - 315	-0.401	-0.082	0.009	5026	-9.05	**<.001**
$F0_{maxloc}$	45 - 241	0.49	0.1	0.009	5026	11.06	**<.001**
$F0_{maxloc}$	45 - 315	-0.642	-0.131	0.009	5026	-14.5	**<.001**
$F0_{maxloc}$	241 - 315	-1.132	-0.231	0.009	5026	-25.58	**<.001**

Table C.7 Multiple comparisons of memory load × tone types by Mandarin participants with Tukey adjustments. Only the results of comparisons between the same tone types are shown. Effect sizes are shown, using Cohen's *d.* Significant findings (p <.05) are shown in bold.

Features	Thai tone	*d*	Estimate	*SE*	*df*	*t*	*p*
$F0_{mean}$	T21	-0.167	-0.01	0.005	138	-2.17	0.479
$F0_{mean}$	T33	-0.04	-0.003	0.005	138	-0.53	1
$F0_{mean}$	T45	0.364	0.023	0.005	138	4.75	**<.001**
$F0_{mean}$	T241	0.042	0.003	0.005	138	0.54	1
$F0_{mean}$	T315	0.33	0.021	0.005	137	4.32	**0.001**
$F0_{excursion}$	T21	0.072	0.01	0.024	38	0.42	1
$F0_{excursion}$	T33	0.421	0.06	0.024	38	2.49	0.308
$F0_{excursion}$	T45	0.231	0.033	0.024	38	1.36	0.931
$F0_{excursion}$	T241	0.225	0.032	0.024	38	1.33	0.941
$F0_{excursion}$	T315	0.522	0.075	0.024	38	3.08	0.095
$F0_{maxloc}$	T21	0.051	0.01	0.021	60	0.49	1
$F0_{maxloc}$	T33	0.151	0.031	0.021	60	1.46	0.901
$F0_{maxloc}$	T45	-0.123	-0.025	0.021	60	-1.19	0.971
$F0_{maxloc}$	T241	0.045	0.009	0.021	60	0.43	1
$F0_{maxloc}$	T315	0.106	0.022	0.021	60	1.03	0.989

Table C.8 Multiple comparisons of talker variability× tone type by Mandarin participants with Tukey adjustments. Only the results of comparisons between the same tone types are shown. Effects size are shown, using Cohen's *d.* Significant findings (p <.05) are shown in bold.

Features	Thai tones	*d*	Estimate	*SE*	*df*	*t*	*p*
Duration	T21	-0.372	-0.043	0.007	5026	-5.94	**<.001**
Duration	T33	0.054	0.006	0.007	5026	0.87	0.997
Duration	T45	0.213	0.025	0.007	5026	3.39	**0.024**
Duration	T241	-0.019	-0.002	0.007	5026	-0.3	1
Duration	T315	-0.012	-0.001	0.007	5026	-0.2	1
$F0_{mean}$	T21	0.159	0.01	0.004	5026	2.54	0.249
$F0_{mean}$	T33	0.062	0.004	0.004	5026	0.99	0.993
$F0_{mean}$	T45	0.043	0.003	0.004	5026	0.69	1
$F0_{mean}$	T241	-0.356	-0.022	0.004	5026	-5.67	**<.001**
$F0_{mean}$	T315	0.011	0.001	0.004	5026	0.18	1
$F0_{excursion}$	T21	-0.136	-0.02	0.009	5026	-2.18	0.474
$F0_{excursion}$	T33	-0.043	-0.006	0.009	5026	-0.69	1
$F0_{excursion}$	T45	-0.003	0	0.009	5026	-0.04	1
$F0_{excursion}$	T241	-0.012	-0.002	0.009	5026	-0.19	1
$F0_{excursion}$	T315	0.063	0.009	0.009	5026	1	0.992
$F0_{maxloc}$	T21	-0.171	-0.035	0.013	5026	-2.73	0.163
$F0_{maxloc}$	T33	0.617	0.126	0.013	5026	9.84	**<.001**
$F0_{maxloc}$	T45	-0.024	-0.005	0.013	5026	-0.38	1
$F0_{maxloc}$	T241	0.201	0.041	0.013	5026	3.21	**0.044**
$F0_{maxloc}$	T315	0.9	0.183	0.013	5026	14.4	**<.001**

Multiple comparisons of fixed factors for Vietnamese data

Table C.9 LMER model results for Vietnamese listener imitation in terms of duration.

Effects	Estimate	*SE*	*t*
(Intercept)	-0.0202	0.01	-2.001
LowMemoryLoad	-0.0158	0.01	-1.115
ConstantTalker	0.0492	0.01	3.466
ConstantVowel	0.0097	0.01	0.681
Tone33	0.1085	0.01	7.623
Tone45	0.0515	0.01	3.617
Tone241	0.0843	0.01	5.936
Tone315	-0.0220	0.01	-1.543
LowMemoryLoad:ConstantTalker	-0.0029	0.02	-0.144
LowMemoryLoad:ConstantVowel	-0.0030	0.02	-0.149
ConstantTalker:ConstantVowel	-0.0062	0.02	-0.309
LowMemoryLoad:Tone33	-0.0191	0.02	-0.951
LowMemoryLoad:Tone45	0.0134	0.02	0.670
LowMemoryLoad:Tone241	-0.0017	0.02	-0.087
LowMemoryLoad:Tone315	0.0298	0.02	1.486
ConstantTalker:Tone33	-0.0835	0.02	-4.163
ConstantTalker:Tone45	-0.0723	0.02	-3.592
ConstantTalker:Tone241	-0.0472	0.02	-2.356
ConstantTalker:Tone315	-0.0579	0.02	-2.881
ConstantVowel:Tone33	-0.0411	0.02	-2.050
ConstantVowel:Tone45	-0.0279	0.02	-1.390
ConstantVowel:Tone241	-0.0257	0.02	-1.278
ConstantVowel:Tone315	-0.0192	0.02	-0.956
LowMemoryLoad:ConstantTalker:ConstantVowel	-0.0108	0.03	-0.381
LowMemoryLoad:ConstantTalker:Tone33	0.0500	0.03	1.768
LowMemoryLoad:ConstantTalker:Tone45	0.0139	0.03	0.490
LowMemoryLoad:ConstantTalker:Tone241	0.0188	0.03	0.663

LowMemoryLoad:ConstantTalker:Tone315	0.0203	0.03	0.717
LowMemoryLoad:ConstantVowel:Tone33	0.0438	0.03	1.546
LowMemoryLoad:ConstantVowel:Tone45	0.0233	0.03	0.825
LowMemoryLoad:ConstantVowel:Tone241	0.0160	0.03	0.565
LowMemoryLoad:ConstantVowel:Tone315	0.0147	0.03	0.519
ConstantTalker:ConstantVowel:Tone33	0.0401	0.03	1.414
ConstantTalker:ConstantVowel:Tone45	0.0200	0.03	0.706
ConstantTalker:ConstantVowel:Tone241	0.0064	0.03	0.225
ConstantTalker:ConstantVowel:Tone315	0.0084	0.03	0.295
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone33	-0.0501	0.04	-1.252
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone45	-0.0153	0.04	-0.383
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone241	-0.0045	0.04	-0.112
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone315	0.0101	0.04	0.251

Table C.10 LMER model results for Vietnamese listener imitation in terms of $F0_{mean}$.

Effects	Estimate	*SE*	*t*
(Intercept)	0.0052	0.01	0.894
LowMemoryLoad	0.0025	0.01	0.309
ConstantTalker	-0.0147	0.01	-1.916
ConstantVowel	0.0005	0.01	0.060
Tone33	-0.0333	0.01	-4.336
Tone45	-0.0529	0.01	-6.896
Tone241	-0.0382	0.01	-4.992
Tone315	0.0472	0.01	6.144
LowMemoryLoad:ConstantTalker	0.0022	0.01	0.200
LowMemoryLoad:ConstantVowel	-0.0060	0.01	-0.558
ConstantTalker:ConstantVowel	0.0013	0.01	0.121
LowMemoryLoad:Tone33	0.0132	0.01	1.225
LowMemoryLoad:Tone45	0.0091	0.01	0.838
LowMemoryLoad:Tone241	-0.0053	0.01	-0.492
LowMemoryLoad:Tone315	-0.0090	0.01	-0.836
ConstantTalker:Tone33	0.0015	0.01	0.141
ConstantTalker:Tone45	0.0165	0.01	1.517
ConstantTalker:Tone241	0.0384	0.01	3.554
ConstantTalker:Tone315	0.0070	0.01	0.643
ConstantVowel:Tone33	0.0005	0.01	0.050
ConstantVowel:Tone45	0.0132	0.01	1.219
ConstantVowel:Tone241	0.0074	0.01	0.680
ConstantVowel:Tone315	0.0054	0.01	0.503
LowMemoryLoad:ConstantTalker:ConstantVowel	0.0079	0.02	0.516
LowMemoryLoad:ConstantTalker:Tone33	0.0092	0.02	0.606
LowMemoryLoad:ConstantTalker:Tone45	0.0021	0.02	0.139
LowMemoryLoad:ConstantTalker:Tone241	0.0046	0.02	0.302
LowMemoryLoad:ConstantTalker:Tone315	-0.0094	0.02	-0.614

LowMemoryLoad:ConstantVowel:Tone33	-0.0017	0.02	-0.110
LowMemoryLoad:ConstantVowel:Tone45	-0.0127	0.02	-0.833
LowMemoryLoad:ConstantVowel:Tone241	-0.0018	0.02	-0.115
LowMemoryLoad:ConstantVowel:Tone315	0.0007	0.02	0.047
ConstantTalker:ConstantVowel:Tone33	0.0126	0.02	0.822
ConstantTalker:ConstantVowel:Tone45	-0.0079	0.02	-0.517
ConstantTalker:ConstantVowel:Tone241	-0.0149	0.02	-0.975
ConstantTalker:ConstantVowel:Tone315	0.0094	0.02	0.617
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone33	-0.0239	0.02	-1.106
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone45	0.0063	0.02	0.290
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone241	0.0055	0.02	0.253
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone315	-0.0175	0.02	-0.810

Table C.11 LMER model results for Vietnamese listener imitation in terms of $F0_{excursion}$.

Effects	Estimate	*SE*	*t*
(Intercept)	-0.0051	0.02	-0.323
LowMemoryLoad	-0.0193	0.02	-0.866
ConstantTalker	0.0275	0.02	1.723
ConstantVowel	-0.0005	0.02	-0.033
Tone33	0.1030	0.02	6.442
Tone45	0.0456	0.02	2.854
Tone241	0.0561	0.02	3.515
Tone315	0.2606	0.02	16.262
LowMemoryLoad:ConstantTalker	0.0024	0.02	0.106
LowMemoryLoad:ConstantVowel	0.0242	0.02	1.073
ConstantTalker:ConstantVowel	-0.0103	0.02	-0.456
LowMemoryLoad:Tone33	0.0051	0.02	0.226
LowMemoryLoad:Tone45	0.0307	0.02	1.363
LowMemoryLoad:Tone241	-0.0293	0.02	-1.302
LowMemoryLoad:Tone315	0.0163	0.02	0.722
ConstantTalker:Tone33	-0.0159	0.02	-0.706
ConstantTalker:Tone45	-0.0118	0.02	-0.523
ConstantTalker:Tone241	-0.0247	0.02	-1.097
ConstantTalker:Tone315	-0.0590	0.02	-2.611
ConstantVowel:Tone33	-0.0010	0.02	-0.045
ConstantVowel:Tone45	0.0196	0.02	0.872
ConstantVowel:Tone241	-0.0084	0.02	-0.374
ConstantVowel:Tone315	0.0259	0.02	1.147
LowMemoryLoad:ConstantTalker:ConstantVowel	-0.0134	0.03	-0.421
LowMemoryLoad:ConstantTalker:Tone33	-0.0060	0.03	-0.189
LowMemoryLoad:ConstantTalker:Tone45	-0.0094	0.03	-0.294
LowMemoryLoad:ConstantTalker:Tone241	-0.0003	0.03	-0.010
LowMemoryLoad:ConstantTalker:Tone315	0.0213	0.03	0.670

LowMemoryLoad:ConstantVowel:Tone33	-0.0220	0.03	-0.690
LowMemoryLoad:ConstantVowel:Tone45	-0.0329	0.03	-1.034
LowMemoryLoad:ConstantVowel:Tone241	-0.0191	0.03	-0.599
LowMemoryLoad:ConstantVowel:Tone315	-0.0604	0.03	-1.896
ConstantTalker:ConstantVowel:Tone33	0.0149	0.03	0.469
ConstantTalker:ConstantVowel:Tone45	-0.0135	0.03	-0.424
ConstantTalker:ConstantVowel:Tone241	0.0142	0.03	0.444
ConstantTalker:ConstantVowel:Tone315	-0.0148	0.03	-0.465
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone33	-0.0018	0.04	-0.040
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone45	0.0353	0.05	0.783
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone241	0.0141	0.05	0.313
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone315	0.0325	0.05	0.722

Table C.12 LMER model results for Vietnamese listener imitation in terms of $F0_{maxloc}$.

Effects	Estimate	*SE*	*t*
(Intercept)	0.0015	0.02	0.065
LowMemoryLoad	-0.0142	0.03	-0.438
ConstantTalker	0.0258	0.02	1.046
ConstantVowel	0.0258	0.02	1.046
Tone33	0.0082	0.02	0.331
Tone45	-0.0816	0.02	-3.299
Tone241	-0.1128	0.02	-4.569
Tone315	0.2019	0.02	8.143
LowMemoryLoad:ConstantTalker	-0.0073	0.03	-0.209
LowMemoryLoad:ConstantVowel	-0.0307	0.03	-0.882
ConstantTalker:ConstantVowel	-0.0291	0.03	-0.834
LowMemoryLoad:Tone33	0.0221	0.03	0.634
LowMemoryLoad:Tone45	-0.0131	0.03	-0.376
LowMemoryLoad:Tone241	0.0005	0.03	0.015
LowMemoryLoad:Tone315	0.0042	0.03	0.120
ConstantTalker:Tone33	-0.1921	0.03	-5.507
ConstantTalker:Tone45	-0.0178	0.03	-0.509
ConstantTalker:Tone241	-0.1027	0.03	-2.946
ConstantTalker:Tone315	-0.2545	0.03	-7.289
ConstantVowel:Tone33	-0.0629	0.03	-1.804
ConstantVowel:Tone45	-0.0385	0.03	-1.105
ConstantVowel:Tone241	-0.0158	0.03	-0.453
ConstantVowel:Tone315	-0.0554	0.03	-1.586
LowMemoryLoad:ConstantTalker:ConstantVowel	0.0252	0.05	0.512
LowMemoryLoad:ConstantTalker:Tone33	0.0183	0.05	0.371
LowMemoryLoad:ConstantTalker:Tone45	0.0188	0.05	0.381
LowMemoryLoad:ConstantTalker:Tone241	0.0299	0.05	0.607
LowMemoryLoad:ConstantTalker:Tone315	0.0172	0.05	0.350

LowMemoryLoad:ConstantVowel:Tone33	0.0307	0.05	0.623
LowMemoryLoad:ConstantVowel:Tone45	0.0317	0.05	0.644
LowMemoryLoad:ConstantVowel:Tone241	0.0526	0.05	1.067
LowMemoryLoad:ConstantVowel:Tone315	0.0593	0.05	1.204
ConstantTalker:ConstantVowel:Tone33	0.1026	0.05	2.082
ConstantTalker:ConstantVowel:Tone45	0.0181	0.05	0.367
ConstantTalker:ConstantVowel:Tone241	0.0316	0.05	0.641
ConstantTalker:ConstantVowel:Tone315	0.0472	0.05	0.957
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone33	-0.0411	0.07	-0.591
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone45	-0.0260	0.07	-0.373
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone241	-0.0514	0.07	-0.738
LowMemoryLoad:ConstantTalker:ConstantVowel:Tone315	-0.0366	0.07	-0.525

Table C.13 Multiple comparisons of tone main effect on imitations by Vietnamese participants with Tukey adjustments. Effect sizes are shown, using Cohen's *d.* Significant findings (p <.05) are shown in bold.

Features	Thai tones	*d*	Estimates	*SE*	*df*	*t*	*p*
Duration	21 - 33	-0.565	-0.064	0.005	5019	-12.76	**<.001**
Duration	21 - 45	-0.182	-0.021	0.005	5019	-4.1	**<.001**
Duration	21 - 241	-0.502	-0.057	0.005	5019	-11.32	**<.001**
Duration	21 - 315	0.297	0.034	0.005	5019	6.69	**<.001**
Duration	33 - 45	0.384	0.043	0.005	5019	8.66	**<.001**
Duration	33 - 241	0.063	0.007	0.005	5019	1.42	0.614
Duration	33 - 315	0.862	0.097	0.005	5019	19.46	**<.001**
Duration	45 - 241	-0.321	-0.036	0.005	5019	-7.23	**<.001**
Duration	45 - 315	0.478	0.054	0.005	5019	10.79	**<.001**
Duration	241 - 315	0.799	0.09	0.005	5019	18.01	**<.001**
$F0_{mean}$	21 - 33	0.387	0.024	0.003	5019	8.73	**<.001**
$F0_{mean}$	21 - 45	0.614	0.037	0.003	5019	13.86	**<.001**
$F0_{mean}$	21 - 241	0.334	0.02	0.003	5019	7.53	**<.001**
$F0_{mean}$	21 - 315	-0.771	-0.047	0.003	5019	-17.38	**<.001**
$F0_{mean}$	33 - 45	0.227	0.014	0.003	5019	5.13	**<.001**
$F0_{mean}$	33 - 241	-0.053	-0.003	0.003	5019	-1.2	0.754
$F0_{mean}$	33 - 315	-1.158	-0.07	0.003	5019	-26.13	**<.001**
$F0_{mean}$	45 - 241	-0.28	-0.017	0.003	5019	-6.32	**<.001**
$F0_{mean}$	45 - 315	-1.385	-0.084	0.003	5019	-31.24	**<.001**
$F0_{mean}$	241 - 315	-1.105	-0.067	0.003	5019	-24.89	**<.001**
$F0_{excursion}$	21 - 33	-0.738	-0.094	0.006	5019	-16.65	**<.001**
$F0_{excursion}$	21 - 45	-0.436	-0.055	0.006	5019	-9.84	**<.001**
$F0_{excursion}$	21 - 241	-0.2	-0.025	0.006	5019	-4.5	**<.001**
$F0_{excursion}$	21 - 315	-1.913	-0.243	0.006	5019	-43.13	**<.001**
$F0_{excursion}$	33 - 45	0.301	0.038	0.006	5019	6.81	**<.001**
$F0_{excursion}$	33 - 241	0.538	0.068	0.006	5019	12.14	**<.001**
$F0_{excursion}$	33 - 315	-1.175	-0.149	0.006	5019	-26.53	**<.001**
$F0_{excursion}$	45 - 241	0.237	0.03	0.006	5019	5.33	**<.001**
$F0_{excursion}$	45 - 315	-1.477	-0.187	0.006	5019	-33.31	**<.001**
$F0_{excursion}$	241 - 315	-1.713	-0.217	0.006	5019	-38.61	**<.001**
$F0_{maxloc}$	21 - 33	0.385	0.076	0.009	5019	8.68	**<.001**
$F0_{maxloc}$	21 - 45	0.522	0.102	0.009	5019	11.77	**<.001**
$F0_{maxloc}$	21 - 241	0.762	0.15	0.009	5019	17.18	**<.001**
$F0_{maxloc}$	21 - 315	-0.384	-0.075	0.009	5019	-8.65	**<.001**

$F0_{maxloc}$	33 - 45	0.137	0.027	0.009	5019	3.09	**0.017**
$F0_{maxloc}$	33 - 241	0.378	0.074	0.009	5019	8.52	**<.001**
$F0_{maxloc}$	33 - 315	-0.768	-0.151	0.009	5019	-17.34	**<.001**
$F0_{maxloc}$	45 - 241	0.241	0.047	0.009	5019	5.43	**<.001**
$F0_{maxloc}$	45 - 315	-0.905	-0.178	0.009	5019	-20.43	**<.001**
$F0_{maxloc}$	241 - 315	-1.146	-0.225	0.009	5019	-25.83	**<.001**

Table C.14 Multiple comparisons of memory load × tone types by Vietnamese participants with Tukey adjustments. Only the results of comparisons between the same tone types are shown. Effect sizes are shown, using Cohen's *d.* Significant findings (p <.05) are shown in bold.

Features	contrast	*d*	*Estimate*	*SE*	*df*	*t*	*p*
duration	T21	0.19	0.021	0.007	685	3.03	0.075
duration	T33	0.055	0.006	0.007	681	0.87	0.997
duration	T45	-0.06	-0.007	0.007	683	-0.95	0.995
duration	T241	0.062	0.007	0.007	687	0.98	0.993
duration	T315	-0.251	-0.028	0.007	685	-4	**0.003**
$F0_{mean}$	T21	-0.042	-0.003	0.005	139	-0.55	1
$F0_{mean}$	T33	-0.223	-0.014	0.005	138	-2.92	0.109
$F0_{mean}$	T45	-0.13	-0.008	0.005	139	-1.69	0.798
$F0_{mean}$	T241	0	0	0.005	139	0	1
$F0_{mean}$	T315	0.25	0.015	0.005	139	3.26	**0.044**
$F0_{excursion}$	T21	0.074	0.009	0.017	43	0.54	1
$F0_{excursion}$	T33	0.147	0.019	0.017	43	1.07	0.985
$F0_{excursion}$	T45	-0.071	-0.009	0.017	43	-0.52	1
$F0_{excursion}$	T241	0.353	0.045	0.017	43	2.57	0.262
$F0_{excursion}$	T315	0.035	0.004	0.017	43	0.26	1
$F0_{maxloc}$	T21	0.137	0.027	0.024	47	1.11	0.982
$F0_{maxloc}$	T33	-0.048	-0.009	0.024	47	-0.39	1
$F0_{maxloc}$	T45	0.108	0.021	0.024	47	0.88	0.997
$F0_{maxloc}$	T241	-0.01	-0.002	0.024	48	-0.08	1
$F0_{maxloc}$	T315	-0.033	-0.006	0.024	47	-0.26	1

Table C.15 Multiple comparisons of talker variability × tone type by Vietnamese participants with Tukey adjustments. Only the results of comparisons between the same tone types are shown. Effect sizes are shown, using Cohen's *d.* Significant findings (p <.05) are shown in bold.

Features	contrast	*d*	estimate	*SE*	*df*	*t*	*p*
duration	T21	-0.372	-0.042	0.007	5020	-5.93	**<.001**
duration	T33	0.080	0.009	0.007	5019	1.28	0.959
duration	T45	0.152	0.017	0.007	5019	2.43	0.311
duration	T241	-0.055	-0.006	0.007	5019	-0.88	0.997
duration	T315	-0.009	-0.001	0.007	5019	-0.14	1.000
$F0_{mean}$	T21	0.180	0.011	0.004	5019	2.87	0.114
$F0_{mean}$	T33	0.074	0.005	0.004	5019	1.18	0.975
$F0_{mean}$	T45	-0.068	-0.004	0.004	5019	-1.09	0.985
$F0_{mean}$	T241	-0.388	-0.024	0.004	5019	-6.19	**<.001**
$F0_{mean}$	T315	0.137	0.008	0.004	5019	2.19	0.467
$F0_{excursion}$	T21	-0.159	-0.02	0.008	5019	-2.54	0.248
$F0_{excursion}$	T33	-0.065	-0.008	0.008	5019	-1.04	0.989
$F0_{excursion}$	T45	-0.045	-0.006	0.008	5019	-0.72	0.999
$F0_{excursion}$	T241	-0.047	-0.006	0.008	5019	-0.74	0.999
$F0_{excursion}$	T315	0.216	0.027	0.008	5019	3.44	**0.021**
$F0_{maxloc}$	T21	-0.071	-0.014	0.012	5019	-1.13	0.981
$F0_{maxloc}$	T33	0.652	0.128	0.012	5019	10.41	**<.001**
$F0_{maxloc}$	T45	-0.041	-0.008	0.012	5019	-0.66	1.000
$F0_{maxloc}$	T241	0.361	0.071	0.012	5019	5.74	**<.001**
$F0_{maxloc}$	T315	1.107	0.217	0.012	5019	17.66	**<.001**

Table C.16 Multiple comparisons of tone types across language groups with Tukey adjustments. Only the results of comparisons between the same tone types are shown. Effect sizes are shown, using Cohen's *d.* Significant findings (p <.05) are shown in bold.

Features	Thai tones	*d*	Estimate	*SE*	*df*	*t*	*p*
Duration	T21	-0.091	-0.010	0.005	1370	-2.05	0.566
Duration	T33	0.025	0.003	0.005	1367	0.55	1.000
Duration	T45	-0.089	-0.010	0.005	1369	-2.01	0.590
Duration	T241	0.039	0.004	0.005	1373	0.87	0.997
Duration	T315	0.121	0.014	0.005	1365	2.74	0.158
$F0_{mean}$	T21	-0.001	0.000	0.003	277	-0.02	1.000
$F0_{mean}$	T33	0.203	0.013	0.003	276	3.75	**0.008**
$F0_{mean}$	T45	0.270	0.017	0.003	277	4.98	**<.001**
$F0_{mean}$	T241	-0.259	-0.016	0.003	277	-4.77	**<.001**
$F0_{mean}$	T315	-0.010	-0.001	0.003	276	-0.18	1.000
$F0_{excursion}$	T21	0.381	0.052	0.015	79	3.45	**0.029**
$F0_{excursion}$	T33	0.146	0.020	0.015	79	1.32	0.946
$F0_{excursion}$	T45	0.162	0.022	0.015	79	1.46	0.902
$F0_{excursion}$	T241	0.556	0.075	0.015	79	5.03	**<.001**
$F0_{excursion}$	T315	-0.541	-0.073	0.015	79	-4.90	**<.001**
$F0_{maxloc}$	T21	0.073	0.015	0.016	105	0.91	0.996
$F0_{maxloc}$	T33	0.252	0.050	0.016	105	3.13	0.065
$F0_{maxloc}$	T45	0.141	0.028	0.016	105	1.75	0.762
$F0_{maxloc}$	T241	-0.122	-0.024	0.016	105	-1.52	0.883
$F0_{maxloc}$	T315	-0.095	-0.019	0.016	105	-1.18	0.974

Appendix D Participant information sheet and the consent form

Participant Information Sheet – General (Unspecified)

Project Title: Cognitive Factors in Perception and Imitation of Thai Tones by Mandarin versus Vietnamese Speakers

Project Summary:

You are invited to participate in a research study being conducted by Juqiang CHEN (PhD student at the MARCS institute, WSU), under the Supervision of Catherine Best (primary supervisor), Mark Antoniou, Benjawan Kasisopa. The research is about how you perceive and produce tones. You will perform three tasks.

How is the study being paid for?

This research is funded by the Research Training Scheme (RTS) of the MARCS institutes, WSU.

What will I be asked to do?

You will be asked to

1. Listen to some Thai tones and categorise them into your native tone system (task 1)
2. Listen to some pairs of Thai tones and decide whether they are the same or not (task 2)
3. Listen to some Thai tone and imitate them as accurately as possible. Read the same syllable in your native tones. (task 3)

How much of my time will I need to give?

1. One and half hour for task 1 and task 2 (Session 1)
2. Half an hour for task 3 (Session 2)

What benefits will I, and/or the broader community, receive for participating?

You will be paid 50 dollars or credits (equivalent to 2 hours) for your participation.

Will the study involve any risk or discomfort for me? If so, what will be done to rectify it?

There is no risk or discomfort.

How do you intend to publish or disseminate the results?

It is anticipated that the results of this research project will be published and/or presented in a variety of forums. In any publication and/or presentation, information will be provided in such a way that the participant cannot be identified, except with your permission.

Will the data and information that I have provided be disposed of?

Your data will be used as per Western Sydney University's Open Access Policy. The data collected from the study will be stored securely electronically and in paper files at the MARCS Institute for Brain, Behaviour and Development, Western Sydney University for 5 years. This electronic version of the data will be stored in online research servers and as a database that will be available to researchers, audiologists, and speech pathologists to be used academically upon request and it will be used strictly for academic purposes.

Can I withdraw from the study?

Participation is entirely voluntary and you are not obliged to be involved. If you do participate you can withdraw at any time without giving reason.

If you do choose to withdraw, any information that you have supplied will be deleted. If you want to withdraw from the research, you can e-mail Juqiang Chen at 19057910@student.westernsydney.edu.au.

Can I tell other people about the study?

Yes, you can tell other people about the study by providing them with the Chief Investigator's (Juqiang Chen) contact details. They can contact the Chief Investigator to discuss their participation in the research project and obtain a copy of the information sheet.

What if I require further information?

Please contact *Juqiang Chen* should you wish to discuss the research further before deciding whether or not to participate

Juqiang Chen, PhD student, +61 420 286 368, 19057910@student.westernsydney.edu.au

Catherine Best, primary supervisor, +61 2 9772 6760

What if I have a complaint?

If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through Research Engagement, Development and Innovation (REDI) on Tel +61 2 4736 0229 or email humanethics@westernsydney.edu.au

Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

If you agree to participate in this study, you may be asked to sign the Participant Consent Form. The information sheet is for you to keep and the consent form is retained by the researcher/s.

This study has been approved by the Western Sydney University Human Research Ethics Committee. The Approval number is *[H12560]*.

Consent Form – General (Unspecified)

Project Title: Cognitive Factors in Perception and Imitation of Thai Tones by Mandarin versus Vietnamese Speakers

I hereby consent to participate in the above named research project.

I acknowledge that:

- I have read the participant information sheet (or where appropriate, have had it read to me) and have been given the opportunity to discuss the information and my involvement in the project with the researcher/s
- The procedures required for the project and the time involved have been explained to me, and any questions I have about the project have been answered to my satisfaction.

I consent to:

☐ *Having my responses to perceptual experiment recorded*

☐ *Having my imitation of tones audio recorded*

Data publication, reuse and storage

This project seeks consent for the data provided to be used in any other projects in the future.

To make reuse of the data possible it will be stored under Western Sydney University's Open Access Policy.

I understand that in relation to publication of the data **my involvement is confidential and the information gained during the study may be published but no information about me will be used in any way that reveals my identity.**

the researchers intend to make the non-identified data from this project available for other research projects

☐ **I can withdraw from the study at any time without affecting my relationship with the researcher/s, and any organisations involved, now or in the future.**

Signed:

Name:

Date:

This study has been approved by the Human Research Ethics Committee at Western Sydney University. The ethics reference number is: H*12560*

What if I have a complaint?

If you have any complaints or reservations about the ethical conduct of this research, you may contact the Ethics Committee through Research Engagement, Development and Innovation (REDI) on Tel +61 2 4736 0229 or email humanethics@westernsydney.edu.au.

Any issues you raise will be treated in confidence and investigated fully, and you will be informed of the outcome.

Appendix E Chen, J., Best, C. T., Antoniou, M., & Kasisopa, B. (2019). Cognitive factors in perception of Thai tones by naïve Mandarin listeners. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 1684–1688). Australasian Speech Science and Technology Association Inc.

Cognitive factors in perception of Thai tones by naïve Mandarin listeners

Juqiang Chen[1], Catherine T. Best[1,2], Mark Antoniou[1], Benjawan Kasisopa[1]

[1]The MARCS Institute, Western Sydney University, Australia
[2]Haskins Laboratories, New Haven CT, USA
J.Chen2/C.Best/M.Antoniou/B.Kasisopa@westernsydney.edu.au

ABSTRACT

Memory load and task-irrelevant phonetic variations influence discrimination of non-native segmental contrasts. We tested how these factors modulate perceptual assimilation and/or discrimination of non-native lexical tone contrasts, relative to Perceptual Assimilation Model (PAM) [1-2] predictions. When perceptually assimilating Thai tones to their native tone system, Mandarin listeners showed sensitivity to native allophonic differences only if memory load was low, but were unaffected by phonetic variations in talkers and vowels. However, AX discrimination decreased with either talker or vowel variability. Unlike non-native segment perception, where discrimination is poorer under high memory load than lower load, tone discrimination was not diminished by high load (long interstimulus interval). PAM-driven predictions were supported across the cognitive manipulations: when two Thai tones were categorized into two native categories (TC) they were better discriminated than when one or both Thai tones were uncategorized (UC/UU). Overlapping choices for TC assimilations can reduce discrimination accuracy.

Keywords: cross-language perception, cognitive load, lexical tones, Thai tones, Mandarin listeners

1. INTRODUCTION

Perceptual attunement to native speech constrains perception of non-native contrasts. This influence from one's native language is theorized by the Perceptual Assimilation Model (PAM) [1] to occur by way of perceptual assimilation. A given non-native phone may be perceptually assimilated to the native phonological system in one of three ways: (1) as Categorized to a native phoneme; (2) as an Uncategorized phone that falls between native phonemes; or (3) as a Non-Assimilable (NA) non-speech sound. Consequently, PAM claims that discrimination is better if two non-native phones are assimilated into two native categories (Two Category assimilation: TC) than if they are both assimilated into the same native category but differ in their discrepancy from the native "ideal" (Category Goodness difference: CG), which is in turn better than if they are equally good/poor exemplars of one category (Single Category: SC).

Naïve listeners can categorize non-native consonants and vowels [3]–[5] as well as lexical tones [7]–[9] into their native categories, and their perceptual assimilation patterns can predict their performance in discriminating non-native contrasts. However, those studies did not systematically manipulate cognitive factors. A number of studies have identified a distinction between a phonetic mode and a phonemic/phonological mode in speech perception [6]–[8], as a function of cognitive load.

Memory load, the capacity to hold a rapidly decaying memory for limited time [9], can cause a switch between modes. In discrimination tasks it is operationalized by manipulating inter-stimulus intervals (ISI). With long ISIs (1500 ms; high load), English listeners discriminated the Hindi retroflex-dental stop contrast according to L1 phonology which has only an alveolar stop, whereas with short ISIs (500 ms; low memory load) they discriminated phonetic level differences. Similarly, German listeners' discrimination of Japanese segmental length contrasts was negatively affected by high memory load (ISI = 2500ms) [10].

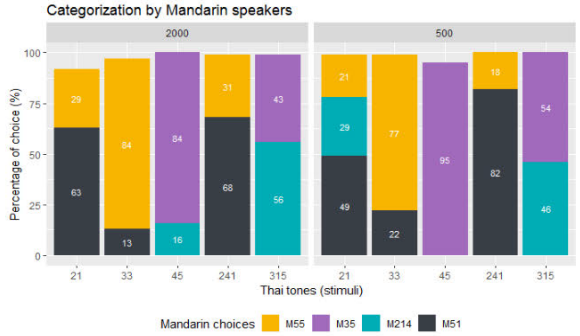
Another cognitive factor explored in previous discrimination studies is attentional control, the ability to allocate attention between task-relevant and irrelevant information [11]. In [10], German listeners discrimination of Japanese consonant length was adversely affected when task-irrelevant information (pitch variation) was added to the task.

Talker variability also leads to attention shifts and increased cognitive load. One theory is that talker variability leads listeners to form multiple phonetic interpretations for a particular acoustic pattern, holding the alternatives in working memory while shifting attention to evaluate them [12]. This suggests that accommodating talker variability demands more working memory resources.

Studies on cognitive factors in cross-language perception have mostly been restricted to consonants and vowels, whereas few cross-language tone per-

rized into a native Mandarin tone category, we set the categorization criterion to 70% of responses [17]. Both T33 and T45 were Categorized in all cognitive conditions; T21 and T315 were Uncategorized, differing from [13]. T241 was Categorized in the short response interval condition but Uncategorized in the long response interval condition.

Figure 1: Categorization of Thai tones by Mandarin listeners in the two response interval conditions.



3. EXPERIMENT 2: DISCRIMINATION

3.1 Method

3.1.1 Participants and Stimulus Materials were the same as in Experiment 1.

3.1.2 Procedure

An AX task ("same-different") was used because it allows better control of ISI; it was used in many previous discrimination studies [10]. ISI (500ms vs. 2000ms) was a between-subjects factor, as we reasoned that listeners may not be able to switch between phonetic and phonological mode within an experiment. Eight blocks (talker × vowel variability) were randomized for each participant.

3.1.2 Data analysis

In order to minimize decision bias, we calculated *d'* for discrimination performance. For each tone pair in each cognitive condition, *d'* scores were calculated using the formula *d'* = Z (hit rate) - Z (false positive rate) with adjustments made for probabilities of 0 (=.01) and 1 (=.99). Hit is defined as the number of correct responses ("different" responses on AB or BA trials). False positive is defined as the number of incorrect responses ("different" responses on AA or BB trials).

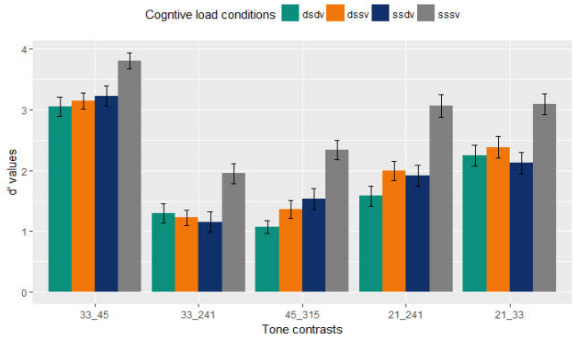
3.2 Results

17847 raw data points were collected (with 73 missing points removed). The *d'* scores for each participant were calculated separately for each tone pair, and for each block, yielding 40 data points for each participant. We fitted the data using a Linear Mixed Effect Regression (LMER) model with *d'* as the dependent variable, and ISI, talker variability, vowel variability and tone pairs as fixed factors, and subject as the random intercept.

To calculate the *p*-values for the fixed effects, we used the Kenward-Roger approximation to the degrees of freedom, as recommend by [18], and the *Anova* function from the *car* package in R, with test specified as "F". The main effect of ISI and all interactions involving ISI were non-significant. However, there were significant main effects of talker variability, $F(1, 1054) = 55.99$, $p < .001$, vowel variability, $F(1, 1054) = 63.80$, $p < .001$ and tone pair, $F(1, 1054) = 109.89$, $p < .001$.

In addition there was a significant interaction between speaker and vowel variability, $F(1, 1054) = 27.67$, $p < .001$. As ISI did not affect the perception, we plotted mean and standard error bars only in terms of vowel and talker variability in Figure 2.

Figure 2: Discrimination of Thai tone contrasts by Mandarin listeners in different cognitive load conditions[a].



[a]Notes: dsdv stands for different speaker different vowel block; dssv stands for different speakers same vowel block; ssdv stands for same speaker different vowels; sssv stands for same speaker same vowel.

Moreover, we did multiple comparisons to test effects of different cognitive factors with the R-package *lsmeans*. Significant differences were found between the same-talker-same-vowel block and the different-talker-different-vowel block, the cognitively easiest versus the most difficult blocks, respectively, β = -1.00, SE = .091, $t(1054)$ = -10.93, p <.001. In addition, when talkers were the same, there was a significant difference between same vowel and different vowels, β = -.86, SE = .091, $t(1054)$ = -9.37 , $p < .001$. Similarly, when the vowel was the same, talker variability had a significant difference, β = -.82, SE = .91, $t(1054)$ = - 9.01 , $p < .001$. Other combinations of talker and vowel conditions were not significantly different.

To test PAM-driven predictions of tone pair contrasts, we did multiple comparisons (Table 1).

Table 1: Multiple comparisons of tone contrasts with PAM-driven predictions

Contrast types	Tone contrasts		t	p
TC_TC/UC	33_45	33_241	18.52	**
TC_UC	33_45	45_315	16.87	**
TC_UC/UU	33_45	21_241	11.37	**
TC_UC	33_45	21_33	8.23	**
TC/UC_UC	33_241	45_315	-1.65	.467
TC/UC_UC/UU	33_241	21_241	-7.15	**
TC/UC_UC	33_241	21_33	10.29	**
UC_UC/UU	45_315	21_241	-5.50	**
UC_UC	45_315	21_33	-8.64	**
UC/UU_UC	21_241	21_33	-3.14	*

Note: * indicates < .05; ** < .001

4. GENERAL DISCUSSION

Generally, we found that different cognitive factors affect discrimination and categorization of non-native tones. First, the response interval effects reflect the influence of memory load in categorization task. The initial information listeners get from speech is low-low level phonetic information. For example, in low memory load, they chose M214 as a response for T21, based on the phonetic similarity between T21 and the allotone of M214 (M21) [13]. But when they were required to wait before selecting their choice answer, the phonetic details faded and all they retained was the more abstract, categorical phonological information. Thus they chose phonologically more similar falling tone M51 for T21, ignoring their phonetic differences. Additionally, phonological processing takes more time than phonetic processing; thus the response interval effect could also reflect two different levels of processing.

Neither vowel nor talker variability affected categorization. Vowel and talker variability in categorization task existed in blocks, not in each trial. Thus, when listeners focused on one tone at a time in each block, the distraction of the talker and vowel variation may have been less than in discrimination. Moreover, the categorization task is more phonological in nature, especially in long response time condition where phonetic details decay and listeners have more time for high-level processing and thus in this case their choices were less susceptible to low level phonetic variations. This supports the argument of phonological constancy that perceivers can assimilate indexical properties of unfamiliar talkers into the key indexical features of their native speech community [19]. Thus they were immune to talker and vowel variation within each block.

Conversely, both vowel and talker variability affected discrimination. AX task involves more phonetic than phonological processing (using more bottom-up stimulus information) and thus is more susceptible to phonetic variation than categorization task. However, ISI did not lead to different performances in tone discrimination, unlike consonant discrimination, where in long ISI listeners fail to distinguish differences that they can do in short ISI. The reason may be that the acoustic cues for consonants are short in duration, and thus are more likely to decay in short-term memory. Tones are longer in duration (in this study extending over the whole syllable). Thus they are less susceptible to decay in short-term memory.

Most PAM-motivated predictions work in different cognitive conditions. TC contrast is better discriminated than the UC and UU contrasts as predicted by PAM. Within UC contrasts, 21_33 (UC) was better discriminated than 45_315 (UC). This could be because 45_315 has a stronger overlapping response choices (M35), leading to more confusion. However, 33_241 (TC/UC) was perceived significantly worse than 21_241(UC/UU) and 21_33 (UC). This could be because both T33 and 241 were assimilated to complementarily overlapping native choices (both yielded M1 and M4 choices), similar to what has been found in unassimilated vowel pairs [20].

5. CONCLUSION

Cross-language tone categorization and discrimination were each affected by different cognitive factors. The longer response interval may have led to decay of low-level phonetic information in the categorization task, shifting listeners to rely more on phonological similarity between native and non-native tones. Categorization was not affected by talker and vowel variability, however, indicating that listeners were to maintain phonological constancy. Showing the opposite pattern, tone discrimination was robust across long and short ISIs, unlike prior findings on perception of non-native segments, but it was affected by the low-level task-irrelevant phonetic variations of talker and vowel variability. PAM-driven predictions (TC>UC>UU) were largely upheld under the different cognitive load conditions. Another novel finding was that overlapping native category choices even for TC assimilation types can decrease discrimination performance on the affected non-native tone pairs. This study has implications for theories of speech perception in general and in particular for other models of non-native and second language speech perception such as the Speech Learning Model (SLM) [21] and the Second Language Speech Perception model (L2LP) [22].

ception studies have investigated cognitive factors in discrimination. Those that have did not use theory-driven predictions about L1 influences. In addition, memory load and attention control were most often manipulated in discrimination tasks, which generally involve more low-level phonetic processing than categorization tasks. Thus it is unknown whether these cognitive factors affect cross-language perceptual assimilation similarly.

Our study manipulated memory load and attention control in non-native tone categorization and discrimination tasks. Memory load was operationalized as ISI in the discrimination task and as response interval (time between the end of the stimulus and the signal to select an L1 category) in the categorization task. Attention control was operationalized by manipulating talker and vowel variability.

Discrimination was tested first in each of two sessions to minimize effects of prior categorization on performance. Due to the multiple cognitive load conditions, it was not feasible to test all Thai tone pairs. Based on a previous study using the same stimuli [13], we selected three Thai tone contrasts that met the required PAM assimilation patterns: T33-T45 (TC), T315-T45 (SC), T33-T241 (UC). After the first session, participants were asked to come back two weeks later for a second session in which we tested another two pairs: T21-T241 (UC), T21-T33 (TC), to compare with a Vietnamese group in a larger project.

Based on PAM, we predicted that Mandarin listeners would discriminate T33-T45 (TC) and T21-T33 (TC) better than T33-T241 (UC) and T21-T241 (UC). T315-T45 (SC) should be the most difficult contrast to distinguish. In order to evaluate these predictions, we report the categorization experiment before the discrimination.

2. EXPERIMENT 1: CATEGORIZATION

2.1. Method

2.1.1. Participants

28 native speakers of Mandarin participated in both experiments 1 and 2, divided into two groups for each response interval/ISI condition (ISI_{short}: M_{age} = 24 years, SD = 4; 8 females; ISI_{long}: M_{age} = 25 years, SD = 6; 10 females). Participants completed a background questionnaire before the test. All had normal hearing and none had experience with Thai or more than two years of formal musical training.

2.1.2 Stimulus materials

Two syllables (/ma/, /mi/) were chosen because they are real words for each native tone in both Thai and Mandarin. The target Thai syllables were each read several times by two female native Thai speakers. These informants had no experience with other tone languages. Two tokens of each target item that were judged to be correct and most natural-sounding to a third native Thai speaker were selected.

We used Chao values [14] to provide a priori phonetic descriptions of the tones in each language. In Chao notation, F0 height at tone onset and offset is referenced by numbers 1-5 ranging from low to high. Thai, the target language, has three level tones (characterized as high-level T45, mid-level T33, low-level T21) and two contour tones (rising T315 and falling T241) [15]. Mandarin has four tones: a level tone M55; a rising tone M35; a falling-rising tone M214; and a falling tone M51 [16].

Response interval condition (2000ms vs. 500ms) was a between-subjects factor, while talker variability (same vs. different) and vowel variability (same vs. different vowels in each trial: /ma/, /mi/) were blocked within each group.

2.1.3 Procedure

Participants were tested individually in the testing booth (at Western Sydney University, or UNSW). Stimuli were presented on a Dell Latitude 7280 laptop running E-Prime Professional 2. Stimuli were presented via Sennheiser HD 280 Pro headphones at 72 dB SPL.

Before the test session, participants completed 10 practice trials. The categorization task had 140 trials in total. On each trial, the stimulus token was presented and listeners made a forced-choice categorization judgment to their native tones (four Pinyin options) via a key press as quickly as possible within a 3s timeout. They then heard the tone again and rated how well it fitted their chosen native category on a 7-point scale. (1 = poor, 7 = perfect, 4 = OK).

2.2. Results

3897 data points were collected (23 missing points removed). We fitted the data with a multinomial regression model. The full model was built with Mandarin tone choice as a dependent measure, and response interval, talker variability, vowel variability and Thai tones as fixed factors. Fixed factors were subtracted one at a time and compared to the full model to determine the effect of cognitive load. Both response interval ($\chi^2(12) = 92.37$, $p < .0001$) and Thai tone ($\chi^2(12) = 5630$, $p < .0001$) showed significant effects in Likelihood ratio tests. However, talker variability and vowel variability were not significant. Mean percent selection of each Mandarin L1 tone for each Thai tone are plotted in Figure 1.

To determine whether a Thai tone was catego-

6. REFERENCES

[1] C. T. Best, "A direct realist view of cross-language speech perception.," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.

[2] C. T. Best and M. D. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language Learning & Language Teaching*, vol. 17, O.-S. Bohn and M. J. Munro, Eds. Amsterdam: John Benjamins Publishing Company, 2007, pp. 13–34.

[3] C. T. Best and W. Strange, "Effects of Phonological and Phonetic Factors on Cross-Language Perception of Approximants," *J. Phon.*, vol. 20, no. 3, pp. 305–330, Jul. 1992.

[4] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *J. Acoust. Soc. Am.*, vol. 109, no. 2, pp. 775–794, Feb. 2001.

[5] M. D. Tyler, C. T. Best, A. Faber, and A. G. Levitt, "Perceptual assimilation and discrimination of non-native vowel contrasts," *Phonetica*, vol. 71, no. 1, pp. 4–21, 2014.

[6] J. F. Werker and R. C. Tees, "Phonemic and phonetic factors in adult cross-language speech perception," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1866–1878, Jun. 1984.

[7] D. B. Pisoni, "Auditory and phonetic memory codes in the discrimination of consonants and vowels," *Percept. Psychophys.*, vol. 13, no. 2, pp. 253–260, Jun. 1973.

[8] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Percept. Psychophys.*, vol. 37, no. 1, pp. 35–44, Jan. 1985.

[9] A. Baddeley and B. A. Wilson, "Prose recall and amnesia: implications for the structure of working memory," *Neuropsychologia*, vol. 40, no. 10, pp. 1737–1743, 2002.

[10] Y. Asano, "Discriminating Non-Native Segmental Length Contrasts Under Increased Task Demands," *Lang. Speech*, pp. 1–21, Oct. 2017.

[11] T. Isaacs and P. Trofimovich, "Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech," *Appl. Psycholinguist.*, vol. 32, no. 1, pp. 113–140, Jan. 2011.

[12] H. Nusbaum and T. M. Morin, "Paying Attention to Differences Among Talkers," in *Speech perception, production and linguistic structure*, Y. Tohkura, Y. Sagisaka, and E. Vatikiotis-Bateson, Eds. Tokyo: Ohmasha Publishing, 1992, pp. 113–134.

[13] J. Chen, C. T. Best, M. Antoniou, and B. Kasisopa, "Cross-language categorisation of monosyllabic Thai tones by Mandarin and Vietnamese speakers: L1 phonological and phonetic influences," presented at the Proceedings of the Seventeenth Australasian International Conference on Speech Science and Technology, 2018, pp. 168–172.

[14] Chao. Y.R., "A system of tone-letters," *Maitre Phon.*, vol. 45, pp. 24–27, 1930.

[15] A. Reid *et al.*, "Perceptual assimilation of lexical tone: The roles of language experience and visual information," *Atten. Percept. Psychophys.*, vol. 77, no. 2, pp. 571–591, Feb. 2015.

[16] M. Yip, *Tone*. Cambridge: Cambridge University Press, 2002.

[17] M. Antoniou, M. D. Tyler, and C. T. Best, "Two ways to listen: Do L2-dominant bilinguals perceive stop voicing according to language mode?," *J. Phon.*, vol. 40, no. 4, pp. 582–594, Jul. 2012.

[18] U. Halekoh and S. Hojsgaard, "A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest," *J. Stat. Softw.*, vol. 59, no. 9, pp. 1–30, 2014.

[19] C. T. Best, "Devil or Angel in the Details?: Perceiving phonetic variation as information about phonological structure," in *Phonetics-Phonology Interface: Representations and Methodologies*, J. Romero and M. Riera, Eds. 2015, pp. 3–31.

[20] M. M. Faris, C. T. Best, and M. D. Tyler, "Discrimination of uncategorised non-native vowel contrasts is modulated by perceived overlap with native phonological categories," *J. Phon.*, vol. 70, pp. 1–19, Sep. 2018.

[21] J. E. Flege, "Second-language speech learning: Theory, findings, and problems," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. 1995, pp. 229–273.

[22] P. Escudero, *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. The Netherlands: Utrecht: LOT, 2005.

Appendix F Chen, J., Best, C. T., & Antoniou, M. (2019). Cognitive Factors in Thai-Naïve Mandarin Speakers' Imitation of Thai Lexical Tones. *Proc. Interspeech 2019*, 2653–2657. https://doi.org/10.21437/Interspeech.2019-1403

Cognitive factors in Thai-naïve Mandarin speakers' imitation of Thai lexical tones

Juqiang Chen[1], Catherine T. Best[1,2], Mark Antoniou[1],

[1]The MARCS Institute, Western Sydney University, Australia
[2]Haskins Laboratories, New Haven CT, USA
J.Chen2/C.Best/M.Antoniou@westernsydney.edu.au

Abstract

The present study investigated how cognitive factors, memory load and attention control, affected imitation of Thai tones by Mandarin speakers with no prior Thai experience. Mandarin speakers lengthened the syllable duration, enlarged the F0 excursion and moved some F0 max location earlier compared with the stimuli, even in the immediate imitation condition. Talker variability had a larger impact on imitation than memory load, whereas vowel variability did not have any effect. Perceptual assimilation patterns partially influenced imitation performance, suggesting phonological categorization in imitation and a perception-production link.

Index Terms: lexical tones, imitation, cognitive factors, non-native speech contrasts

1. Introduction

In order to learn to speak a second language, language learners often rely on imitating the words produced by native speakers. Imitation links speech perception and production in a natural way, providing an excellent opportunity to examine the link between perceiving and producing a non-native language without the need of orthographic knowledge.

It is argued that imitation is constrained by native phonological categories. Native speakers fail to imitate phonetic details of vowel or VOT continua [1-2]. Instead their imitations reflect their native phonological categories. Spanish monolinguals, Spanish speakers of English and English monolinguals were asked to imitate a stop consonant voice onset time (VOT) continuum ranging from /da/ to /ta/ in [2]. Their imitations did not show a linear incremental increase in VOT, but instead formed two or three VOT response categories that matched the native phoneme boundaries. Similarly, Finnish children and adults imitated a /æ/ to /ɑ/ vowel continuum, showing categorical imitation patterns relevant to their native phonological categories [1].

In addition, second language learning research has reported that imitation by second language speakers is modulated by their native language phonology. For example, when asked to imitate eight American English vowels (e.g. /i/, /ɪ/, /e/, /ɛ/, /æ/, /ʌ/, /ɑ/, /u/), native Mandarin speakers showed influence of their native language [3]: /ɛ/, /æ/ with no Mandarin counterparts were imitated worse than /i/ and /u/ with counterparts in Mandarin. These findings support well-established second language speech learning theories, such as the Speech Learning Model (SLM) [4] and the Perceptual Assimilation Model (PAM) [5]. Both models claim that attunement to the native language will cause listeners' inability to discern the phonetic differences between pairs of sounds in the L2, or between L2 and L1 sounds via "equivalence classification" (SLM) and "perceptual assimialtion" (PAM). In addition, without accurate perceptual "targets" to guide the sensorimotor learning of L2 sounds, production of the L2 sounds will be inaccurate [4].

However, some researchers argue against imitation as always being phonologically constrained [6]. English speakers identified and imitated Mandarin tones whereas Korean speakers identified and imitated English consonants. Researchers found that imitation was generally more accurate than the identification. Therefore, they inferred that imitation was not always constrained by native phonology as in identification and that L2 imitation may bypass some aspects of phonological encoding. Furthermore, they proposed that participants can operate on a phonetic mode of processing without any access to native phonology. In [7], native speakers of Polish imitated English unreleased plosives in two-stop sequences in two imitation conditions. In one condition, they imitated the stimuli immediately and, in another condition, they read a digit after the auditory input before imitating. Performance was native-like in the first condition and it was significantly impeded in the second. This suggests that native language constraints on imitation of non-native phones is modulated by the mode of imitation. According to the Automatic Selective Perception (ASP) model, the change between a phonetic and phonological mode in perception is modulated by cognitive load [8]. In this paper, we extend ASP to imitation to make predictions about how cognitive factors may influence imitation performance.

Very few studies have investigated non-native tone imitation by tone language speakers. Even fewer tone imitation studies have considered cognitive factors. Therefore, the present study examines how cognitive load factors affect the imitation of Thai tones by Mandarin-native speakers with no prior experience with Thai. Thai and Mandarin differ in the number of tones and types of tones in their native inventories. We used Chao values [9] to provide a priori phonetic description of the tones in each language. In Chao notation, F0 height at tone onset and offset is referenced by numbers 1-5 ranging from low to high. Thai has three level tones (characterized as high-level T45, mid-level T33, low-level T21) and two contour tones (rising T315 and falling T241) [10]. Mandarin has four tones: a level tone M55; a rising tone M35; a falling-rising tone M214; and a falling tone M51[11]. Pervious perception work [12] has indicated that Thai-naïve Mandarin listeners assimilated T45 and T315 into a single Mandarin tone category, M35. T33 and T21 were categorized as M55 and M214 respectively. T241 was split between M55 and M51, thus uncategorized.

Two cognitive factors, namely memory load and attention control demand, were systematically manipulated in the study. Memory load is the capacity to hold a rapidly decaying

memory for a limited period of time [13]. It can affect imitation because the auditory memory used in the phonetic mode decays quicker than the memory of more abstract, "encoded" phonological categories. Thus the longer participants are asked to hold the tone in memory, the greater the auditory memory decays, and the more they will have to rely on their longer-lasting phonological memory for imitation. Attention control is the capacity to efficiently allocate attention between task-relevant and irrelevant information [14]. The more complex the stimuli, the higher the demand on attention control, which can affect imitation because participants have to allocate more cognitive resources to process the stimuli to extract tone-related information.

We hypothesize that imitation will be more accurate (i.e., less deviant from the stimulus) when memory load is low because the phonetic information is still available in short memory. Moreover, when acoustic complexity of the stimuli within a block is low, e.g. from one speaker and of one vowel, imitation will be better because participants can attend to tone-related phonetic details. Furthermore, the speaker's native language will interact with the cognitive factors, specifically, since participants in this study are naïve listeners, they have no L2 phonological system to use, imitation will be more constrained by L1 phonology and more L1 accented when memory load is high and stimuli within one block are acoustically complex.

2. Experiment

2.1. Method

2.1.1. Participants

28 native speakers of Mandarin participated in the experiments, divided into two groups for each memory load condition (low load: M_{age} = 24 years, SD = 4; 8 females; high load: M_{age} = 25 years, SD = 6; 10 females). Participants completed a background questionnaire before the test. All had normal hearing and none had experience with Thai or more than two years of formal musical training because musical training can facilitate tone perception and imitation [15].

2.1.2. Stimulus materials

Two syllables (/ma/, /mi/) were chosen for the target stimuli because they are real words in Thai and Mandarin. The target Thai syllables were each recorded several times as produced by two female native Thai speakers who had no experience with any other tone languages. Two tokens of each target item judged to be correct and natural-sounding to a third Thai speaker were used in the imitation study.

2.1.3. Procedure

Memory load was operationalized as the time between the end of the stimuli and the signal for participants to imitate (imitation interval). In the low memory load condition, a message "Imitate now!" was shown 500 ms from the offset of the stimulus to let participants start imitating. In the high memory load condition, it was shown 2000 ms after the offset of the stimulus. In both conditions, participants had 3 s (timeout) to imitate and the inter-trial interval is 1 s.

Attention control was operationalized as within-block talker variability (same vs. different) and vowel variability (same vs. different vowels: /ma/, /mi/). Participants were instructed to imitate stimuli as faithfully as possible after they heard the auditory stimulus. Before the test session, participants completed 10 practice trials. Each participant had 160 trials (2 syllables × 5 tones × 8 conditions × 2 repeats) in total.

Participants were tested individually in testing booths (at Western Sydney University, or UNSW). Stimuli were presented on a Dell Latitude 7280 laptop running E-Prime Professional 2 via Sennheiser HD 280 Pro headphones at 72 dB SPL. Participants' responses were recorded with a portable digital speech recorder (ZOOM H4n) with 41 kHz sampling rate and 16-bit stereo format.

2.1.4. Data analysis

Average pitch, direction, length, extreme point and slope have been reported to be the primary factors affecting the perception of lexical tones [16]. While tone contour direction and slope can be compared via several statistical modelling methods, such as growth curve analysis [17], generalized additive mixed modeling [18] and functional data analysis [19], due to the limited space, this paper focuses on comparing discrete measures that capture features like average pitch, length and extreme points. ProsodyPro [20], a Praat script, was used to extract several acoustic measures from the pitch contours of the Thai stimuli and their imitations: syllable duration, F0 mean, time-normalized 10 points of F0, and F0max location (relative to the syllable duration). F0 maximum to minimum excursion (F0 excursion in short) was calculated by measuring the range of F0 between 10% to 90% of the syllable length (the most stable part). Theoretically, F0 means indicate overall pitch for the three phonologically level tones: T33, T21, T45. F0 excursion could distinguish level tones from contour tones, T241 and T315 which should differ in having different F0max locations. Statistically, in a PCA analysis of lexical tones [21], these acoustic measures outweighed other measures in differentiating Thai, Mandarin, Southern and Northern Vietnamese tones. In order to make F0 means comparable across different speakers, we did Lobanov normalization to F0 means and calculated F0 excursion based on Lobanov-normalized F0 means [22]. It should be noted that the Lobanov-normalized F0 mean values reflect how much an F0 mean for a tone varies from the F0 mean of the speaker.

2.2. Results

2.2.1 Acoustic comparison of Thai tones target stimuli

First, we measured syllable duration, Lobanov-normalized mean F0, F0 excursion, F0max location (in Table 1) to examine how these measures contribute to distinguishing Thai tone target stimuli.

Four linear mixed-effects models were built with the four measures as dependent variables respectively and tone types as the fixed-effects factor and participants and vowels as random-effects factors. To calculate the p-values for the fixed effects (tone types), we used the Kenward-Roger approximation to the degrees of freedom, as recommended by [23], and the *Anova* function from the *car* package in R, with test specified as "F". Significant main effects of tone types as a fixed factor were found for all four measures: syllable duration, $F(4, 33) = 3.10$, $p = .03$; mean F0, $F(4, 33) = 22.59$, $p < .001$, F0max-min excursion size, $F(4, 33) = 12.60$, $p < .001$, F0max location, $F(4, 33) = 72.63$, $p < .001$. This indicates that the selected four measures distinguish the five Thai tones.

Moreover, we conducted multiple comparisons to test how different measures distinguish Thai tones with the R-package

lsmeans. *P*-values smaller than .05 were considered significant. T241 was significantly shorter in syllable duration than T315 and T45, whereas differences in syllable duration among other Thai tones were not significant. As for F0 mean, three phonologically level tones, T33, T21, T45 were distinct from each other. All other tone pairs were significantly different in F0 mean, except for T241-45 and T315-33. T241 had a significantly larger F0 excursion than all other Thai tones whereas T33 had a significantly smaller F0 excursion than other Thai tones. Both T21 and T33 showed no difference in F0 maximum location. T241 had the maximum F0 in the middle of the syllable while both T45 and T315 had it at syllable offset.

Table 1: *Acoustic measures (means) for the Thai target stimuli*[a]

Thai tones	Duration (ms)	F0_mean	F0 excursion	maxF0_loc (%)
T21	597	-0.11	0.19	13
T241	548	0.07	0.27	38
T315	625	-0.02	0.16	98
T33	596	-0.03	0.08	29
T45	612	0.08	0.19	89

[a]Note: F0 mean and F0 excursion are normalized using formula in [22].

2.2.2 Deviation of the imitated Thai tones from the targets

4640 raw imitated tones were collected and 179 were removed because participants started imitating before they had been instructed to do so. To obtain difference scores, we subtracted stimuli data from the imitation data for each acoustic measures: syllable duration, Lobanov-normalized mean F0, F0 excursion, F0max location. The resulting difference scores were selected as dependent variables and each was fitted with a linear mixed-effects model. Memory load (low vs. high), talker variability (same vs. different), vowel variability (same vs. different) and tone types (five Thai targets) were used as fixed factors and participant and imitated vowel (high vowel /i/ and low vowel /a/) were random factors (intercept). Four models were built to test all possible main effects and interactions.

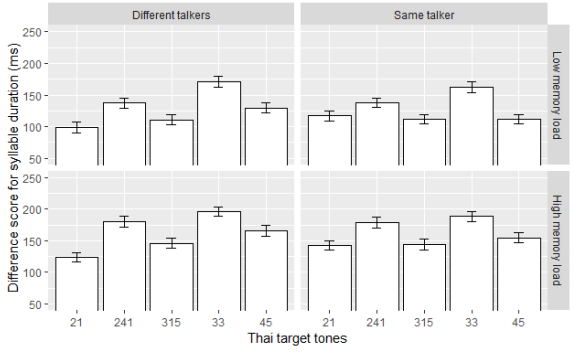


Figure 1: *Difference scores for syllable durations under different cognitive load conditions*

Syllable duration of the imitated Thai tones is shown in Figure 1. We found a significant main effect for tone types, $F(4, 4394) = 69.74$, $p < .001$ and a significant interaction between talker variability and tone types, $F(4, 4394) = 4.58$, $p < .001$. No main effects or interactions were found for vowel variability. We ran multiple comparisons to test the pairwise differences among tone types and the interaction. Positive difference scores indicate that Mandarin speakers lengthened the syllables relative to the original target durations. T33 was significantly lengthened as compared to all other Thai tones. T241 was significantly lengthened as compared to T45, T21, T315, and T45 imitations were significantly longer than the targets for T21 and T315. There was a significant effect of talker variability for T21, $\beta = -18.50$, $SE = 5.71$, $t(4394) = -3.241$, $p = .03$, but not for any other Thai target tones.

Difference scores for Lobanov-normalized F0 mean of the imitated Thai tones are shown in Figure 2. Normalized F0 showed significant main effects of talker variability, $F(1, 4394) = 9.50$, $p = .002$, and tone types, $F(4, 4394) = 54.31$, $p < .001$ and significant interactions between talker variability and tone types, $F(4, 4394) = 11.01$, $p < .001$, and between imitation interval and tone types, $F(4, 4394) = 13.06$, $p < .001$.

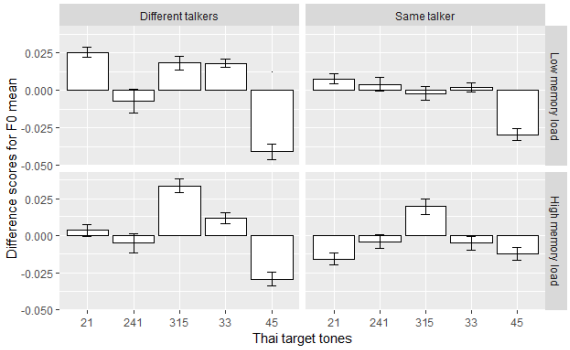


Figure 2: *Difference scores for Lobanov-normalized F0 mean under different cognitive load conditions*

We ran multiple comparisons to test the pairwise differences among tone types and to break down interactions between tone types and memory loads, and between talker variability and tone types. The imitations of both T45 and T315 had significantly larger difference scores compared with other tone imitations. Memory load significantly affected T21, $\beta = 0.02$, $SE = 0.0046$, $t(592) = 4.90$, $p < .001$, T315, $\beta = -0.018$, $SE = 0.0046$, $t(590) = -4.01$, $p = .002$, and marginally significant for T45, $\beta = -0.014$, $SE = 0.0046$, $t(596) = -3.09$, $p = .06$. Talker variability had a significant effect on T21, $\beta = -0.0185$, $SE = 0.0046$, $t(4394) = 3.99$, $p = .0027$, T315, $\beta = -0.017$, $SE = 0.0046$, $t(4394) = 3.70$, $p = .008$, T33, $\beta = -0.016$, $SE = 0.0046$, $t(4394) = 3.57$, $p = .01$ and had a marginal significant effect of T45, $\beta = -0.014$, $SE = 0.0046$, $t(4394) = -3.06$, $p = .06$.

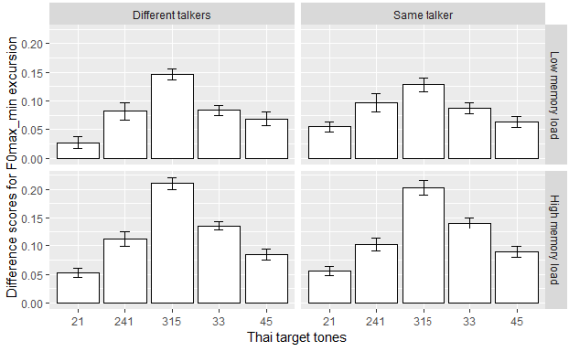


Figure 3: *Difference scores for F0 excursion under different cognitive load conditions.*

For F0 excursion (in Figure 3), we found significant main effects of tone types, $F(4, 4394) = 86.03$, $p < .001$ and a significant interaction between memory load and tone types, $F(4, 4394) = 6.29$, $p < .001$. The difference scores for all tones were positive, suggesting that imitated tones had larger excursion size than the original stimuli. We ran multiple comparisons to test the pairwise differences among tone types and the interaction. T315 had significantly larger difference scores than other tones while T21 had a significantly smaller difference score than other tones. Both T241 and T33 had significantly larger difference scores than T45. Memory load did not significantly affect excursion difference scores for the same tone targets.

F0 max location ratio (Figure 4) showed main effects of talker variability $F(1, 4394) = 7.55$, $p = .006$, and tone types, $F(4, 4394) = 95.46$, $p < .001$, and significant interactions between talker variability and tone types, $F(4, 4394) = 40.95$, $p < .001$ and memory load and tone types, $F(4, 4394) = 5.29$, $p < .001$. The negative difference scores in F0 maximum location indicates earlier position of maximum location in the syllable for imitation than stimuli. We ran multiple comparisons to test the pairwise differences among tone types and the interactions. The difference scores were larger in both T241 and T315 as compared with other tones. Talker variability had significant effects on T21, $\beta = -0.096$, $SE = 0.01614$, $t(4394) = -5.95$, $p < .0001$, T241, $\beta = -0.1224$, $SE = 0.0161$, $t(4394) = -7.58$, $p < .0001$, T33, $\beta = -0.1421$, $SE = 0.0161$, $t(4394) = 8.78$, $p < .0001$. Memory load did not significantly affect F0 max location difference scores for the same tone targets.

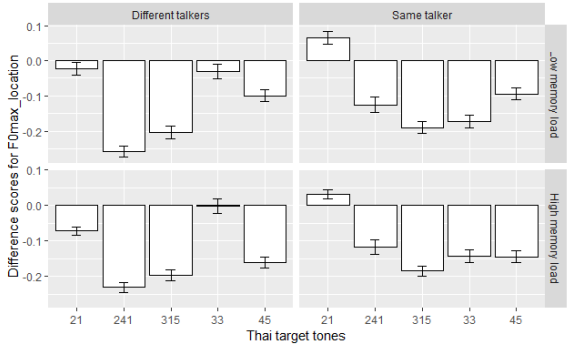


Figure 4: *Difference scores for F0 maximum location ratio under different cognitive load conditions*

3. Discussion

First of all, we found a significant main effect of tone types for all four acoustic difference measures. In other words, the deviation in imitation from the target stimuli in all four measures varied from one tone to another. T21 was the best imitated tone, showing low difference scores in most cases but it was susceptible to memory load and talker variability. The imitations of T241 had significant larger difference scores in syllable duration and F0 excursion and F0max, suggesting that T241 was difficult for Mandarin participants. Given that T241 was not categorized as any Mandarin tone categories [12], the L1 phonological influence is smaller than categorized tones. Thus the observed difficulties in imitation lie more in encoding and producing the Thai targets. T33 was significantly lengthened and was imitated with a larger F0 excursion as compared to the target stimuli. T315 was imitated with higher F0, whereas T45 with lower F0 than the original stimuli. Given that T315 is lower than T45, this means that the two tones were produced with very similar F0 which is in line with the observation that T315 and T45 were both perceptual assimilated into M35 [12]. To sum up, Mandarin speakers lengthened the syllable duration to match the longer stimulus syllables, but they overdid the lengthening. F0 excursion was enlarged in the imitations, suggesting that Mandarin speakers attempted to follow the pitch change of Thai, but over-exaggerated the change. Mandarin speakers tended to delay their F0 peak for T241, T315 and T45, relative to the target stimuli.

Second, some cognitive factors affected imitation performance. Talker variability had significant main effects on two of our four measures, namely F0 mean and F0 max location. It was also involved in interactions with tone types for syllable duration, F0 mean and F0 max location measures. This supports our hypothesis that processing task-irrelevant information increases demands on attention control, leading to poorer performance when imitating.

Memory load did not show a main effect for any of the acoustic measures, but it interacted with tone types for F0 mean, F0 excursion and F0 max location. In multiple comparison tests, for the same tone, memory load modulated F0 mean for T21, T315 and T45. However, for F0 excursion and F0 max location, we did not find any significant effect of memory load for the same tone types. Therefore, memory load did not drastically change imitation performance. The difference scores from the target stimuli in the low memory load condition suggest that participants did not imitate the phonetic details of the target tones correctly. Previous research on imitation of consonant length also showed little effect of memory load [24]. The less than expected effect of memory load could be because when waiting for imitation in the long interval condition, participants rehearsed internally, which reduced the decay of phonetic details perceived from the stimuli, thus diminishing the difference between two memory load conditions. Studies that asked participants to do other tasks while waiting have reported stronger effect of memory load [7].

Vowel variability within a block did not have any main effects or interactions on the imitation of Thai tones for any of the acoustic measures. Unlike talker variability which affects pitch more drastically and requires listeners to adapt to a change in talkers, vowel quality is more intrinsic to lexical tones for tone language speakers. Therefore, processing vowel variability may require less cognitive effort than resolving talker variability.

4. Conclusions

The present study examined how cognitive factors, memory load and talker variability, affected imitation of Thai tones by Mandarin speakers with no experience with Thai. Mandarin speakers lengthened the syllable duration, enlarged the F0 excursion and moved F0 max location earlier, even in the immediate imitation condition. Talker variability affected imitation most and memory load altered imitation performance to a lesser degree. Vowel variability did not have any effect on imitation. Perceptually uncategorized tones are difficult for naïve speakers to imitate and when two tones are assimilated into a single category, the imitation of these two tones resembles each other. These results have implication for theories of non-native speech perception and production (SLM, PAM) as well as pedagogical implications for second language lexical tone training.

5. References

[1] L. Alivuotila, J. Hakokari, J. Savela, R.-P. Happonen, and O. Aaltonen, "Perception and imitation of Finnish open vowels among children, naïve adults, and trained phoneticians," presented at the Proceedings of the 16th International Congress of Phonetic Sciences, 2007, pp. 361–364.

[2] J. E. Flege and W. Eefting, "Imitation of a VOT continuum by native speakers of English and Spanish: evidence for phonetic category formation," *J. Acoust. Soc. Am.*, vol. 83, no. 2, pp. 729–740, Feb. 1988.

[3] G. Jia, W. Strange, Y. Wu, J. Collado, and Q. Guan, "Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure," *The Journal of the Acoustical Society of America*, vol. 119, no. 2, pp. 1118–1130, Jan. 2006.

[4] J. E. Flege, "Second-language speech learning: Theory, findings, and problems," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. 1995, pp. 229–273.

[5] C. T. Best, "A direct realist view of cross-language speech perception.," in *Speech perception and linguistic experience: Issues in cross-language research*, W. Strange, Ed. Timonium, MD: York Press, 1995, pp. 171–204.

[6] Y.-C. Hao and K. de Jong, "Imitation of second language sounds in relation to L2 perception and production," *Journal of Phonetics*, vol. 54, pp. 151–168, Jan. 2016.

[7] A. Rojczyk, A. Porzuczek, and M. Bergier, "Immediate and Distracted Imitation in Second-Language Speech: Unreleased Plosives in English," May 2013.

[8] W. Strange, "Automatic selective perception (ASP) of first and second language speech: A working model," *Journal of Phonetics*, vol. 39, no. 4, pp. 456–466, Oct. 2011.

[9] Chao. Y.R., "A system of tone-letters," *Le Maitre Phonetique*, vol. 45, pp. 24–27, 1930.

[10] A. Reid *et al.*, "Perceptual assimilation of lexical tone: The roles of language experience and visual information," *Atten. Percept. Psychophys.*, vol. 77, no. 2, pp. 571–591, Feb. 2015.

[11] M. Yip, *Tone*. Cambridge: Cambridge University Press, 2002.

[12] J. Chen, C. T. Best, M. Antoniou, and B. Kasisopa, "Cross-language categorisation of monosyllabic Thai tones by Mandarin and Vietnamese speakers: L1 phonological and phonetic influences," presented at the Proceedings of the Seventeenth Australasian International Conference on Speech Science and Technology, 2018, pp. 168–172.

[13] A. Baddeley and B. A. Wilson, "Prose recall and amnesia: implications for the structure of working memory," *Neuropsychologia*, vol. 40, no. 10, pp. 1737–1743, 2002.

[14] T. Isaacs and P. Trofimovich, "Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech," *Applied Psycholinguistics*, vol. 32, no. 1, pp. 113–140, Jan. 2011.

[15] T. L. Gottfried, A. M. Staby, and C. J. Ziemer, "Musical experience and Mandarin tone discrimination and imitation," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2545–2545, Apr. 2004.

[16] J. T. Gandour, "The perception of tone," in *Tone: A linguistic survey*, Academic Press, 1978, pp. 41–76.

[17] P. Tang, I. Yuen, N. X. Rattanasone, L. Gao, and K. Demuth, "Acquisition of weak syllables in tonal languages: acoustic evidence from neutral tone in Mandarin Chinese," *Journal of Child Language*, vol. 46, no. 1, pp. 24–50, Jan. 2019.

[18] M. Wieling, "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English," *Journal of Phonetics*, vol. 70, pp. 86–116, Sep. 2018.

[19] M. Gubian, F. Torreira, and L. Boves, "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts," *Journal of Phonetics*, vol. 49, pp. 16–40, Mar. 2015.

[20] Y. Xu, "ProsodyPro—A tool for large-scale systematic prosody analysis," 2013.

[21] J. Chen, C. T. Best, M. Antoniou, and B. Kasisopa, "Mapping and comparing East and Southeast Asian language tones," presented at the Australia Linguistic Society annual conference, Adelaide, 2018.

[22] B. M. Lobanov, "Classification of Russian Vowels Spoken by Different Speakers," *The Journal of the Acoustical Society of America*, vol. 49, no. 2B, pp. 606–608, Feb. 1971.

[23] U. Halekoh and S. Hojsgaard, "A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest," *Journal of Statistical Software*, vol. 59, no. 9, pp. 1–30, 2014.

[24] Y. Asano and B. Braun, "Does speech production in L2 require access to phonological representations?," presented at the Proceedings of the International Conference on Speech Prosody, 2016, vol. 2016-January, pp. 237–241.