PRODUCTION AND PERCEPTION OF LEXICAL TONE VARIATION IN MANDARIN DIALECTS

Liang Zhao

Doctor of Philosophy

University of York

Language and Linguistic Science

October 2023

ABSTRACT

Mandarin dialects vary in the sound inventory on both the segmental and suprasegmental levels. Previous research has suggested that dialectal variation in the Mandarin language primarily targets the lexical tone system (Chao, 1943; Ho, 2003; Szeto, Ansaldo & Matthews, 2018). Nevertheless, mutual intelligibility has been observed among Mandarin dialects (Tang & van Heuven, 2008, 2009), which leads to the question of how listeners of Mandarin dialects process variability from lexical tone systems. The present dissertation investigates acoustic-phonetic variation across Mandarin dialects with a focus on lexical tone systems and how Standard Mandarin listeners perceptually adapt to unfamiliar lexical tone systems of regional dialects. A speech production study was first conducted for an acoustic-phonetic analysis of the vowel spaces and tone inventories of six Mandarin dialects—Beijing Mandarin, Chengdu Mandarin, Jinan Mandarin, Taiyuan Mandarin, Wuhan Mandarin, and Xi'an Mandarin. The results suggested that Mandarin dialects have comparable segmental systems and disparate tone inventories. Four perception experiments were then conducted in investigating the perceptual mechanisms in processing familiar and unfamiliar Mandarin dialect tone systems and the potential factors that modulate the perception outcome. The results indicated rapid adaptation to the novel tone system within two-minute sentential exposure from the experimental trials using both top-down and bottom-up information. General improvement in accuracy was found as the amount of exposure increased; the post-exposure improvement was greater for tone systems with more dissimilar phonetic contours. Through these findings, I argue for integrated bottom-up and top-down mechanisms for processing lexical tone variation. The perceptual system must actively modify the relative contribution of top-down and bottom-up information for lexical access, based on the reliability of each type of information and the specific tasks. The findings of this thesis have implications for our current understanding of speech perception, especially on the lesser-studied perception of lexical tone variation.

ACKNOWLEDGEMENTS

This journey of PhD study has been exciting, challenging and rewarding, along which I have many people to thank.

First, I would like to extend my deepest gratitude to my advisor Eleanor Chodroff for the generous support and invaluable guidance throughout this process. I greatly benefit from the one-on-one tutorials on scripting and statistical analysis, especially in the first few years; I could not thank her enough for training me in the skillsets that have been crucial for me to grow as an independent researcher in the future. Our discussions on the project have always helped me clarify thinking and writing; our meetings and conversations, either formal or informal, have made my time at York most memorable.

I would also like to thank the Thesis Advisory Panel—George Bailey, Sam Hellmuth and Paul Foulkes for the ample support and feedback in each step as the thesis developed and for sharing their expert insights that guided this project to a better shape.

Thanks also go to my collaborators Shayne Sloggett and Philip Harrison in the Department of Language and Linguistic Science at York. Shayne helped in designing the surprisal sentence manipulation that is critical to the experimental design in my perception study. Phil provided valuable advice on implementing the remote audio collection method. It has been a joy working with them and I feel grateful for the opportunities to learn from the brilliant minds.

I would also like to express my sincere gratitude to my examiners Nino Grillo at the University of York and Cong Zhang at the Newcastle University, for providing extremely helpful comments on the thesis and generously offering insights and discussions on the topic of lexical tone production and perception, which inspired my ideas for further work. I very

much enjoyed the VIVA meeting, feeling grateful for them making the entire process so vigorous and meanwhile friendly.

I also thank my friends, Aini Li, Fani Karageorgou, Huan Wei, Xiaoran Niu, Yue Liu, and everyone in the LLS GTA and Stats Group for the joyful gatherings and mental support.

Finally, I thank my family for the trust and support throughout my study, without whom I would not have gotten here.

TABLE OF CONTENTS

ABSTRACT......ii

ACKNOWLEDGEMENTS...... iii

TABLE OF CONTENTS......v

LIST OF TABLES......x

LIST OF FIGURES......xi

DECLARATION......xiv

Chapter 1 Introduction......1

1.1 The lack-of-invariance problem......3

1.2 Lexical access and spoken word recognition......6

1.2.1 Models of lexical access......7

1.2.2 Discussion and consensus......17

1.3 Constructive speech perception: integrated top-down and bottom-up processing? ...20

1.4 Perceptual learning and adaptation......23

1.4.1 Perceptual learning/adaptation as overall improved perception......24

1.4.2 Perceptual learning/adaptation as reshaped phonetic space......26

1.4.3 Potential factors in perceptual learning/adaptation......29

1.5 The terms in Standard Mandarin and Mandarin dialects......32

1.6 Thesis outline and research questions......33

Chapter 2 Production of Mandarin dialects: comparable segments and distinct tones..35

2.1 Introduction......35

2.1.1 Mandarin dialects as a unique dialect group......36

2.1.2 Segmental systems: Standard Mandarin and Mandarin dialects......39

2.1.2.1 Standard Mandarin segmental system......39

2.1.2.2 Current state of knowledge on dialect segmental inventories......42

2.1.3 Lexical tone inventories: Standard Mandarin and Mandarin dialects......45

2.1.3.1 Standard Mandarin tone system..45
2.1.3.2 Mandarin dialect tone systems..46
2.1.3.3 Systematic word–tone mappings across Mandarin dialects 52
2.1.4 Acoustic parameters of lexical tones 53
2.1.5 Quantification of tone system similarity..54
2.1.6 Speech data collection 57
2.1.7 The present chapter...60
2.2 Methods 61
2.2.1 Subjects...61
2.2.2 Materials 62
2.2.3 Procedure 64
2.2.4 Corpus Annotation...67
2.2.4.1 File Naming 67
2.2.4.2 Transcripts 68
2.2.4.3 Forced Alignment 68
2.2.5 Acoustic measurement...69
2.2.5.1 Formant measurement 69
2.2.5.2 F0 measurement..70
2.3 Results and discussion 71
2.3.1 Vowel spaces 72
2.3.2 Phonetic tone inventories..77
2.3.2.1 Data validation...77
2.3.2.2 Dialect-specific tone inventories 78
2.3.2.3 Phonetic similarity of tones with the same contour type...............84
2.3.2.4 Edit distance of tone systems across dialects 87
2.4 Conclusion 89
2.5 Authorship and publication status 90

Chapter 3 Perception of lexical tone variation with incidental exposure 91

3.1 Introduction 91

3.1.1 Top-down and bottom-up processing of speech 92

3.1.2 The effect of quality and quantity of exposure 93

3.2 Experiment 1: Top-down and bottom-up processing of familiar and unfamiliar tone systems..95

3.2.1 Methods 96

3.2.1.1 Participants 97

3.2.1.2 Materials 97

3.2.1.3 Procedure 98

3.2.1.4 Data analysis 99

3.2.2 Results..100

3.2.2.1 Accuracy (Experiment 1)..100

3.2.2.2 Response time (Experiment 1)..102

3.2.3 Discussion..103

3.3 Experiment 2: Conditions on adaptation to an unfamiliar lexical tone system—the role of quantity and quality of incidental exposure 106

3.3.1 Methods 108

3.3.1.1 Participants 108

3.3.1.2 Materials 108

3.3.1.3 Procedure 109

3.3.1.4 Data analysis 110

3.3.2 Results..111

3.3.2.1 Accuracy (Experiment 2)..111

3.3.2.2 Response time (Experiment 2)..114

3.3.2.3 The effect of minimal pairs..117

3.3.3 Discussion..119

3.4 Conclusion 120

3.5 Authorship and publication status 121

Chapter 4 Perception of lexical tone variation with explicit exposure 122

4.1 Introduction 122

4.1.1 Adaptation with increased exposure..123

4.1.2 Phonetically similar vs. phonologically similar tone system.......................124

4.2 Experiment 3: Adaptation to Chengdu vs. Jinan tones with minimal pairs.............125

4.2.1 Methods 126

4.2.1.1 Participants 126

4.2.1.2 Materials 126

4.2.1.3 Procedure 127

4.2.1.4 Data analysis 128

4.2.2 Results..129

4.2.2.1 Accuracy (Experiment 3)...129

4.2.2.2 Response times (Experiment 3) 131

4.2.3 Discussion...133

4.3 Experiment 4: Adaptation to Chengdu vs. Jinan tones with no minimal pairs.........135

4.3.1 Methods 135

4.3.1.1 Participants 135

4.3.1.2 Materials 136

4.3.1.3 Procedure 136

4.3.1.4 Data analysis 136

4.3.2 Results..137

4.3.2.1 Accuracy (Experiment 4)...137

4.3.2.2 Response time (Experiment 4)...139

4.3.3 Discussion...141

4.4 Combined analysis...145

4.4.1 Data analysis...145

4.4.2 Results.....145

4.4.2.1 Accuracy.....145

4.4.2.2 Response time.....147

4.4.3 Discussion.....149

4.5 Conclusion.....150

4.6 Authorship and publication status.....151

Chapter 5 Conclusion.....152

5.1 Summary of the key findings.....152

5.1.1 The production study: comparable segments and distinct phonetic tones.....152

5.1.2 The perception studies: rapid adaptation through integrative perception.....153

5.2 Further discussions.....155

5.2.1 The presence of top-down influence in lexical access.....156

5.2.2 The relative contribution of bottom-up and top-down information: modifiable given reliability.....158

5.2.3 Perceptual adaptation as remapping between tone category and phonetic contour.....160

5.2.4 The effect of increased exposure: potential ceiling effect?.....160

5.3 Limitations and directions for future research.....161

APPENDIX.....164

a. List of monosyllabic words.....164

b. List of disyllabic words.....166

c. List of surprisal sentence pairs.....167

d. The North Wind and the Sun.....170

e. Wo Chun poem.....171

f. The initial model results in Experiment 2 (Section 3.2).....172

g. The follow-up model results with the tone factors in Experiment 2 (Section 3.3).....174

REFERENCES.....176

VITA.....202

LIST OF TABLES

Table 1.1 Summary of the debatable aspects of the reviewed lexical access models. The sign “-” means unspecified or unclear. 18

Table 2.1 The consonant inventory of Standard Mandarin (and Mandarin dialects). 40

Table 2.2 Pinyin representation of the consonant inventory of Standard Mandarin (and Mandarin dialects). 41

Table 2.3 The vowel inventory of Standard Mandarin (left) and the Pinyin representation (right). 42

Table 2.4 The tone system of Standard Mandarin: examples with the syllable /ma/. 45

Table 2.5 Tone systems of six Mandarin dialects in Chao tone notation according to the two references published in 2002 (Ref. 1) and 1989 (Ref. 2). Bolded numbers correspond to a discrepancy between the two sources....................................47

Table 2.6 Intelligibility percentage scores for Beijing listeners presented with the six dialects. 55

Table 2.7 Example of non-normalised and normalised edit distance. 57

Table 2.8 Participants in the speech production experiment (F for female, M for male). 62

Table 2.9 Mean and standard deviation of F1 (Hz), F2 (Hz) and duration (ms) for /i, ə, a, u/ in each of the six dialects...75

Table 2.10 Comparison of tone systems from different sources. Asterisks mark the tones that show completely different contour patterns between measured and recorded tones. 81

Table 2.11 Updated Chao tone numerals based on measured acoustics (see also in Table 2.10). 88

Table 2.12 Average tone edit distance of the six dialects relative to Standard Mandarin. 88

Table 3.1 An example sentence item across surprisal conditions..................................98

LIST OF FIGURES

Figure 1.1 Schematic diagram of human lexical access (Stevens, 2008)..........................13

Figure 1.2 Ye & Connine's modified TRACE model from Fig.1 in Ye & Connine (1999), titled as "Phoneme, toneme, and syllable nodes and connections. Larger arrows indicate greater connection strength"..15

Figure 1.3 The Reverse Accessing Model as FIGURE 5 in Gao et al. (2019)..................16

Figure 2.1 Map of Chinese dialect groups based on *Language Atlas of Chinese Dialects* (Wurm et al., 1987) by Wyunhe (2011). The Mandarin group is in light brown. 37

Figure 2.2 Schematic contours of the four lexical tones in Standard Mandarin (right). The figure on the left was from the handbook Xian Dai Han Yu ("Modern Chinese", Wang et al., 2006)..46

Figure 2.3 The measured tone contours of Beijing Mandarin by Shi & Wang (2006). The original figure was Fig. 4 in the paper. Each tone was represented by three lines: the middle stood for the mean values; the upper and the lower lines represented one standard deviation from the mean..49

Figure 2.4 The measured tone contours of Chengdu Mandarin by He (2025). The original figure was Fig. 3 in the paper. Each tone was represented by three lines: the middle stood for the mean values; the upper and the lower lines represented one standard deviation from the mean...50

Figure 2.5 The measured tone contours of Standard Mandarin and Xi'an Mandarin in Liu et al. (2020). The original figure was FIG.1 in the paper. 51

Figure 2.6 Locations of cities where the six Mandarin dialects are spoken (Zhao & Chodroff, 2022)...60

Figure 2.7 Part of the WAV file for the disyllabic word "普通" <pu3 tong1> and its corresponding TextGrid in Praat..69

Figure 2.8 The normalised mean F1 and F2 values by vowel and dialect. 73

Figure 2.9 Scatterplots of normalised F1 and F2 values by vowel in the six dialects. Ellipses indicate two standard errors from the dialect-vowel-specific mean F1 and F2...74

Figure 2.10 Mean vowel durations (ms) across dialects. 75

Figure 2.11 Smoothed tone contours of Standard Mandarin based on the corpus data. 77

Figure 2.12 Mean F0 (semitone) contours of four lexical tones in six dialects. Ribbons currently reflect ± 0.5 standard error from the mean. 78

Figure 2.13 Smoothed mean F0 (Chao tone) contours of four lexical tones in six dialects. Ribbons currently reflect ± 0.5 standard errors from the mean. 79

Figure 2.14 Smoothed Chao tone contours by the four contour patterns (a-d) in the six dialects. 85

Figure 3.1 Smoothed lexical tone contours of Standard Mandarin and Chengdu Mandarin converted to Chao Tone numerals (Zhao & Chodroff, 2022). Ribbons reflect ± 0.2 standard error from the mean. 96

Figure 3.2 Percentage of "correct" responses across dialect and surprisal conditions. "Yes" (plausible; lime green) is treated as correct for low-surprisal conditions, and "no" (not plausible; blue green) for high-surprisal conditions. 101

Figure 3.3 Response times across dialect and surprisal conditions. 103

Figure 3.4 Percentage of correct responses across dialect, surprisal and repetition ("1, 2, 3" refer to the repetition blocks). 112

Figure 3.5 Count of correct responses across dialect, surprisal and tone conditions. 113

Figure 3.6 Response times across dialect and surprisal conditions. 115

Figure 3.7 Response times across dialect, surprisal and repetition conditions ("1, 2, 3" refer to the repetition blocks). 116

Figure 3.8 Response times across dialect, surprisal and tone conditions. 117

Figure 3.9 Response times across dialect, surprisal and presentation conditions in the previous (with-minimal-pair) and the new (no-minimal-pair) experiments. . 118

Figure 4.1 Smoothed lexical tone contours of Standard, Jinan and Chengdu Mandarin converted to Chao Tone numerals (Zhao & Chodroff, 2022). Ribbons reflect ± 0.2 standard error of the mean. 125

Figure 4.2 Percentage of "correct" responses across dialect, exposure and surprisal conditions. "Yes" (plausible; lime green) is treated as the correct response for low-surprisal conditions, and "no" (not plausible; blue green) as correct for high-surprisal conditions. 130

Figure 4.3 Response times across dialect, exposure and surprisal conditions in Experiment 3..132

Figure 4.4 Percentage of "correct" responses across dialect, exposure and surprisal conditions. "Yes" (plausible; lime green) is treated as correct for low-surprisal conditions, and "no" (not plausible; blue green) for high-surprisal conditions. 138

Figure 4.5 Count of "correct" responses across dialect and surprisal conditions by tone categories. 139

Figure 4.6 Response times across dialect, exposure and surprisal conditions. 140

Figure 4.7 Response times across dialect and surprisal conditions by tone categories...141

Figure 4.8 Percentage of "correct" responses across surprisal, dialect, and exposure conditions in the combined analysis. 146

Figure 4.9 Response times across dialect, exposure and surprisal conditions in the combined analysis...148

Figure 4.10 Response times across dialect, surprisal and presentation conditions in the combined analysis...148

DECLARATION

I declare that this dissertation is a presentation of original work and is written following the guidance on the journal-style thesis by the University. Portions of the thesis were published in the papers listed below. I confirm that I am the lead the author of the papers and have major contributions that culminate in the doctoral project. All sources are acknowledged as References.

Zhao, L., Sloggett, S. & Chodroff, E. (2023). Conditions on Adaptation to an Unfamiliar Lexical Tone System: The Role of Quantity and Quality of Exposure. In *Proceedings of the 20th International Congress of Phonetic Science* (ICPhS 2023). Prague, Czech Republic.

Zhao, L. & Chodroff, E. (2022). The ManDi Corpus: A Spoken Corpus of Mandarin Regional Dialects. In *Proceedings of the 13th Conference on Language Resources and Evaluation* (pp. 1985–1990). Marseille, France.

Zhao, L., Sloggett, S., & Chodroff, E. (2022). Top-Down and Bottom-up Processing of Familiar and Unfamiliar Mandarin Dialect Tone Systems. In *Proceedings of the Speech Prosody* (pp. 842–846). Lisbon, Portugal.

Chapter 1 Introduction

Mandarin dialects vary in the phonetic realisation of the speech sounds, especially in the phonetic realisation of the lexical tones. Speakers of Mandarin dialects are no strangers to the notion of "variation" in lexical tones at least in their native dialect. Imagine a household scenario, where the parent is trying to talk sense into a six-year-old and asks for an explanation for the mischievous behaviour by saying "I need you to give me a *jiao1 dai5*" in Beijing Mandarin. The word "*jiao1 dai5*" is used colloquially, meaning an explanation or a solution; the numbers in the Pinyin transcription here critically indicate different lexical tone categories. The six-year-old, who has not learned the word "*jiao1 dai5*", might reply "I don't have any *jiao1 dai4* with me" and look confused. In the kid's view, the word must be "*jiao1 dai4"* which means (adhesive) tape in Beijing Mandarin.

For speakers of the same Mandarin dialect, it seems a minor communicative issue that a tone might change in accordance with the listener's contextual expectations, as this happens occasionally with a particular set of lexical items. However, if the conversation is between speakers of different Mandarin dialects that differ in the phonetic realisation of the entire tone system, the whole utterance would sound dialectal—the segmental sequence might be accented though still familiar, but the exact pitch contours of the tone categories would be unfamiliar. But still, for speakers of different Mandarin dialects, it would likely be the case that an adult Beijing Mandarin speaker would be able to understand the speech of a Chengdu Mandarin speaker and deduce from the context whether the speaker asked for an explanation ("*jiao1 dai5*") or a roll of tape ("*jiao1 dai4*").

In doing so, listeners of Mandarin dialects need to deal with the phonetic variation of the speech sounds, in particular the phonetic variation of the lexical tones. This leads to the questions of how the dialectal speech of Mandarin is perceived and more specifically how listeners deal with variability in the lexical tone system: do they treat unfamiliar tones as unwanted noise, or can they learn from the input and adapt to the novel tone system? While there is a large volume of literature on speech variation, both the speech production and the perception studies on this topic have primarily investigated segmental variation; not much research has been done on how lexical tone systems vary in naturally produced speech and little is known on whether listeners are able to or how they adapt to unfamiliar lexical tone systems.

This dissertation aims to investigate not only the extent of lexical tone variation, but also how it can be processed and potentially adapted to by listeners who are not necessarily familiar with the variation. Specifically, the thesis focuses on the phonetic variation of lexical tones across six Mandarin dialects, seeking to understand the extent of variability in the lexical tone systems (Chapter 2), and how such lexical tone variation is processed by non-native listeners of the dialect, i.e., the perceptual mechanisms being used and the influential factors in adapting to an unfamiliar tone system (Chapter 3 and 4). Speech data of Mandarin dialects is used as the source of tonal variability in the dissertation as they have comparable segmental inventories, but disparate phonetic realisations of the lexical tones, evidenced by the results in Chapter 2.

The following sections in Chapter 1 first introduce rich variability in speech signal, also known as the lack-of-invariance problem, and present some major theoretical claims for speech perception in Section 1.1. Section 1.2 reviews current understanding of how lexical information is extracted from speech signal through examples of lexical access models. Sections 1.3 follows up on a particular debate between interactive and non-interactive processing raised in the previous section and emphasises the assumption of constructive speech perception with more discussion. Section 1.4 overviews the topic of perceptual learning and adaptation with relevant studies and hypotheses in the existing literature. Section 1.5 clarifies the terms related to Standard Mandarin and Mandarin dialects. Section 1.6 outlines the details of each chapter in the thesis and the research questions.

1.1 The lack-of-invariance problem

A central goal in the study of speech perception is to understand the relationship between the speech signal and the intended meaning. However, as the only physical link between speaker and listener, the speech signal is characterised by variability, which may arise from varying phonetic contexts with different adjacent sounds, various acoustic environments such as speech presented in noise, within-speaker differences in terms of speaking rate and emotional state, across-speaker differences such as variation in the speech anatomy, as well as accented and dialectal speech among other sources of variability (Munro & Derwing, 1995; Norris, McQueen & Cutler, 2003; McKay, 2021). For successful speech perception, listeners need to perceive individual speech sounds with varying acoustic properties and researchers seek to understand how they manage to extract linguistic information from highly variable acoustics. However, decades of endeavours to relate acoustic signals to the perceived sound distinctions have reached a widespread recognition that there is *no* consistent one-to-one mapping between the acoustic properties and the linguistic units. Specifically, an acoustic cue can be present in multiple speech sounds and an identifiable speech sound can be represented by multiple acoustic cues. This is often addressed in the existing literature as the lack-of-invariance problem (e.g. Pisoni, 1981; Perkell & Klatt, 1986, 2014; Appelbaum, 1996; Fernández & Cairns, 2011; Hayward, 2014; Raphael, 2021). To put it in a simpler way, there lack consistent acoustic correlates which specifically define distinct speech sounds.

In response to the lack-of-invariance problem, several well-known theories of speech perception hypothesised the existence of certain invariant correlates (e.g. articulatory gestures), mediating between the acoustics and the speech sounds; some argued that the invariant properties can be found directly in the acoustic signal. The following paragraphs summarise the main claims of a few theories that differ in their perspectives on what are the objects of perception, how they are perceived, and whether speech perception involves unique mechanisms specific to speech and to humans. The objects of perception refer to the objects that listeners become directly aware of upon perceiving.

The motor theory of speech perception (see Liberman, Cooper, Shankweiler, & Studdert- Kennedy, 1967; Liberman & Mattingly, 1985) proposes that what listeners perceive from acoustic signals are the abstract intended articulatory gestures (e.g. lip rounding, jaw raising, tongue fronting), rather than the highly variable auditory or acoustic properties. The intended gestures are stored in the brain as invariant commands to the articulators and strongly correspond to the phonological features, while it is unclear how the acoustic properties relate to the intended gestures. The motor theory considers speech perception an innate process and human specific.

The direct realism theory (Fowler, 1986) also argues for articulatory gestures being the objects of perception, but the perceived gestures are not abstract as in the motor theory, but actual vocal tract gestures. Listeners have knowledge of these gestures and how they would interact and result in variance in the acoustic domain such that lack of invariance is no longer a problem. The direct realism approach assumes a non-special function of speech perception in human cognition.

Unlike the motor theory and direct realism, where the invariant objects are found in the articulatory gestures, the theory of acoustic invariance (Stevens & Blumstein, 1981) contends that invariant correlates to phonological features can be extracted directly from the speech signal. Controversial as it remains, the acoustic invariance theory hypothesis strong relationships between acoustic properties and phonological features, and between gestures and phonological features (Nearey, 1995; Hayward, 2014). This theory does not take a strong stance on modularity or innateness.

Concerning the special nature of speech perception, the critical evidence was found in the perceptual differences between listeners' responses to speech and non-speech stimuli, which has underpinned many key aspects of the motor theory. However, shared characteristics in sound perception have also been found between speech and non-speech signals (Stevens & Klatt, 1974; Miller et al., 1976; Pisoni, 1977), and between human and animal species, such as chinchillas, macaques, and Japanese quails (Kuhl & Miller 1975, 1978; Kuhl & Padden, 1982; Kluender, Diehl & Killeen, 1987). For example, Japanese quails showed categorical discrimination of the syllable-initial consonants followed by different vowel segments through training (Kluender et al., 1987). Findings along this line have prompted assumptions and hypotheses of general auditory and learning mechanisms for speech perception (Diehl & Kluender, 1989; Diehl, Lotto & Holt, 2004; Lotto & Holt, 2016). In Diehl et al.'s review (2004), a working hypothesis can be summarised as: perception of speech sounds can be achieved using general auditory and learning mechanisms applicable to sounds in general and potentially across species. It is assumed that extraction of linguistic information from acoustic signals relies on the distributional properties of the stimuli and does not involve gestures (Diehl, Lotto & Holt, 2004). Further to this assumption is the expanding literature on perceptual learning and adaptation, elaborated in Section 1.4 in this chapter.

1.2 Lexical access and spoken word recognition

Research on speech perception broadly focuses on native listeners' ability to recognise and identify speech sounds from the acoustic signal (Beddor, 2017). How lexical information is retrieved for word recognition is commonly associated with a separate area of research, i.e. lexical access (Hayward, 2014) or a more theoretical neutral term, lexical processing (Taft, 2001). Successful retrieval of lexical information relies on the listener's mental lexicon which consists of all known lexical items to the listener in a particular language; they are mentally stored as in the listener's long-term memory. It is generally assumed that two types of information are encoded for each lexical entry—the form and the meaning. The form refers to the phonological or morphological characteristics of the lexical item; the meaning refers to the semantics and syntactic functions (Levelt, 2001; Culter, 2002).

Theoretical frameworks related to lexical access primarily concern what is mentally represented and how the lexical information is retrieved (Taft, 2001). Regarding the nature of mentally stored information, researchers vary in the view on whether the form and the meaning together constitute a lexical entry (Collins & Quillian, 1969), or they are represented on separate levels, and whether the extraction of semantic and syntactic information is indeed part of lexical access or automatically initiated upon word recognition in a post-access period (Norris, 1994; Norris et al., 2003). Proponents of the later views found empirical evidence that memory traces differed between lower-level (morphological) and higher-level (semantic and syntactic) processing (e.g. Craik & Tulving, 1975). With respect to the process of lexical access, it typically involves the activation of the candidates given sensory input, the competition among available candidates, and the selection of one good match to the received input. However, the specific process and the factors that might affect it vary across theories and hypotheses on lexical access.

1.2.1 Models of lexical access

The following section reviews several lexical access models and their major claims about the lexical retrieval process to provide a background for understanding and investigating mechanisms involved in lexical processing, especially lexical processing with tonal representations in the dissertation (Chapter 3 and 4). The models introduced here are those explicitly expounding on the procedures and/or mechanisms for accessing lexical information, from which critical aspects of lexical access are recapped and discussed in the following Section 1.2.2, leading to a key goal of the perception studies in the current thesis. Computational models such as jTRACE (Strauss, Harris & Magnuson, 2007; Bramlett & Wiener, 2022) and TISK model (Hannagan, Magnuson & Grainger, 2013; You & Magnuson, 2018) are left out as they are not within the scope of this study[1]. The models initially proposed for visual word recognition are also mentioned as they contain viewpoints adopted or challenged in the models for spoken word recognition.

Forster's autonomous search model (Forster & Chambers, 1973; Forster, 1976, 1981) is one of the visual word recognition models in support of the serial search mechanism (see also Forster & Bednall, 1976; Murray & Forster, 2004). According to the model, a lexical item is accessed via sequential search among all the available candidates until the one best matching the input is found; the associated lexical information is then automatically retrieved from mental lexicon. Forster's model proposes the existence of access files which function as library catalogues to search for the lexicon and offer the pools of candidates specific to the sensory

[1] The distinctive feature theory and the theory of phonological underspecification were initially proposed as phonological theories concerning feature representation, rather than lexical processing; this is why they were excluded in the current section. However, it is worth mentioning that models of lexical access may differ on whether the phonological representation is featurally underspecified or fully specified. A large volume of literature pertains to this topic (e.g. Chomsky & Halle, 1968; Steriade, 1995; Lahiri & Reetz, 2002, 2010; Wheeldon & Waksler, 2004). There are also hypotheses on phonetic underspecification (e.g. Keating, 1996).

modality. The lexical items in each access file are ordered from most to least frequent. The model also assumes that the initial part of the word may provide sufficient information for lexical access; however, the required amount of the initial portion remains controversial (Bard, Shillcock & Altmann, 1988; Field, 2003). A major criticism of this model is that it fails to explain the observed processing of multi-modality information (McGurk & MacDonald, 1976; Storms, 1998; Vroomen & Gelder, 2000; Hulusic, Harvey, Debattista, Tsingos, Walker, Howard & Chalmers, 2012; Xu & Taft, 2015; Taft, 2015; Pisoni & McLennan, 2016).

Unlike the autonomous search model, Morton's logogen model (1969, 1970, 1979) rejects the idea of sequential search and allows parallel processing of lexical information across modalities and activation of multiple lexical candidates. The model incorporates a mediating device, i.e. "logogen", between sensory input and the lexical representation. Logogens are considered as feature-counting devices, and each designate a particular "threshold" for the required amount of input to activate a potential candidate. Upon receiving the input, logogens start counting the number of features based on the linguistic similarities between the received input and the candidate lexical item; once all the input is received, among the logogens that have reached their thresholds, the one having the most shared features is eventually selected (Morton & Patterson, 1998; Field, 2003). The model also proposes an abstract recovery process, where the above-threshold logogens return to a zero-feature level with a longer period of time than those below the threshold (Morton, 1970, 1979). Moreover, the frequency effect interacts with the threshold value: more frequent items have lower thresholds and potentially require less input for lexical access (Besner & Swan, 1982).

The cohort model (Marslen-Wilson, 1978, 1980) particularly targets the process of spoken word recognition and assumes three stages for lexical access—access, selection and integration. For the access stage, multiple candidates matching the initial acoustic input, e.g.

onset segments, are activated in parallel and constitute a cohort. As more acoustic information becomes available, candidates that mismatch the accumulative input are removed from the cohort until a single item is left. The final stage integrates semantic and syntactic information for the selected item. Criticism of the original version of the cohort model argued that the model did not explain the phenomenon where the input signal clearly mismatches the accessed item, but the item is still recognised (Cole, 1973); also, the effect of word frequency was not discussed (Taft, 1986), and the selection process seemed rather passive and lacking in active participation of the candidates, such as competition (Weber & Scharenborg, 2012). In the revised version of the model (Marslen-Wilson, 1987, 1990), activated candidates compete in the selection stage; the one reaches the highest activation level is selected. The relative activation level of a word depends on the word frequency rather than similarity; once the activation level is reached, the lexical item can be accessed despite any mismatch to the received input (Marslen-Wilson, 1990).

As an alternative to the cohort model, the activate and check model (Taft, 1986) was proposed following the empirical findings on non-word recognition. The results indicated a slowdown effect on non-word stimuli when the initial portion of the input matches the legitimate portions of a word (Taft, 1986). However, the needed amount of the initial portion varied between spoken and visual stimuli: for spoken words, initial input up to the onset segments was sufficient for activating possible candidates, while for visual words it had to be up to the syllable. Taft and Hambly (1986) later coined the term, access code, to refer to the phonetic unit that activates parallel candidates. They assumed the same processing mechanisms for visual and spoken word recognition with the only difference in the access code (Taft & Hambly, 1986). However, Bard, Shillcock and Altmann (1988) questioned the nature of the access code as their findings suggested that successful word recognition mostly happened until the offset of the word and the so-called access code might a contain larger portion than previously suggested.

The TRACE model (Elman & McClelland, 1984; McClelland & Elman, 1986) is one of the better-known connectionist models of speech perception. From the connectionist perspective, there exists a neural network which consists of a large number of interconnected units, each functioning as a processing device (McClelland & Elman, 1986; Protopapas, 1999). The connection between units is weighted given the extent to which activation can be transmitted from one unit to another (Protopapas, 1999). The weighed connection can be modified by bottom-up and top-down influences between lower-level and higher-level processing (McClelland & Elman, 1986).

Specifically, in the TRACE model, the interconnected units are distributed on three layers which correspond to "the feature, phoneme and word" respectively. All the units continuously participate in processing the input and each unit represents a working hypothesis of a perceptual object that can feed units across layers in a dynamic and interactive manner. It is assumed that between the inconsistent units of the same layer (e.g. distinct phonemes), processing is done through inhibitory interactions, whereas for the consistent units of different layers (e.g. a word and its component phonemes), it is through excitatory interactions. According to McClelland & Elman (1986), the model is called TRACE because the processing operations of the spoken stimuli leave traces of processed information over all the layers. A particular aspect of this model is that candidates are activated via any part of the input and the one that has an overall good fit to the input relative to the others is recognised. Two version of the model have been developed to solve specific problems in speech processing. Much more than what is simplified here, TRACE I focuses on the recognition of phonemes (Elman &

McClelland, 1984) and TRACE II primarily targets the lexical effects on phoneme recognition (McClelland & Elman, 1986).

In contrast to the TRACE model which incorporates the top-down process in the network, two other connectionist models explicitly claimed that they are entirely bottom-up. The Shortlist model (Norris, 1994) was proposed to solve unsettled problems in the TRACE model. The model specifies two stages of processing. In the first stage, approximately thirty candidates are selected based on the fitness to the input and constitute a shortlist. Competition of the shortlisted candidates then take place through inhibitory links in the interactive-activation framework (Norris, 1994; Besner & Swan, 2012). A recent update to the model is the Shortlist B model which operates by Bayesian principles and drastically differs from the previous one (Norris & McQueen, 2008). Another model in favour of the bottom-up processing is the Merge model. The model assumes that sub-lexical and lexical information is merged in a strict bottom-up process and top-down feedback is not necessary in spoken word recognition (Norris, 1999; Norris, McQueen & Culter, 2000). These models helped in interpreting certain empirical data, especially in the phonemic decision between word and non-word stimuli, but they largely overlooked the learning effect in speech perception and might not work as a unified model to explain perception of both familiar and unfamiliar speech.

Quite different from the connectionist approach, the model of acoustic landmarks and distinctive features (Stevens, 2008) is developed based on the theory of acoustic invariance (Stevens & Blumstein, 1981) introduced in Section 1.1. The model assumes two paths of operations in perception of words in sequence. Acoustic input is first analysed through the bottom-up path in which 1) the acoustic signal is judged as speech or non-speech via peripheral auditory processing; 2) once determined to be speech input, landmarks of acoustics are detected to recognise the types of stimuli as either vowels, consonants, suprasegmentals, etc.; 3) then acoustics parameters and cues of the stimuli are extracted in the vicinity of landmarks; 4) estimated values of landmarks with a measure of confidence are then calculated and output; 5) intended words are accessed through matching estimated feature bundles with those stored in lexicon and this is the last step of bottom-up process. In running speech, the output of bottom-up processing is assumed to be a cohort of words or word sequences (Stevens, 2008). The top-down process follows as a "synthesis and comparison path" where the output of hypothesised words at the end of bottom-up processing becomes the new input into an external synthesis of parameters and landmarks based on top-down information; as a result, more accurate matches could be decided for the initial speech signal. Figure 1.1 excerpts the schematic representation of Stevens's model (2008).

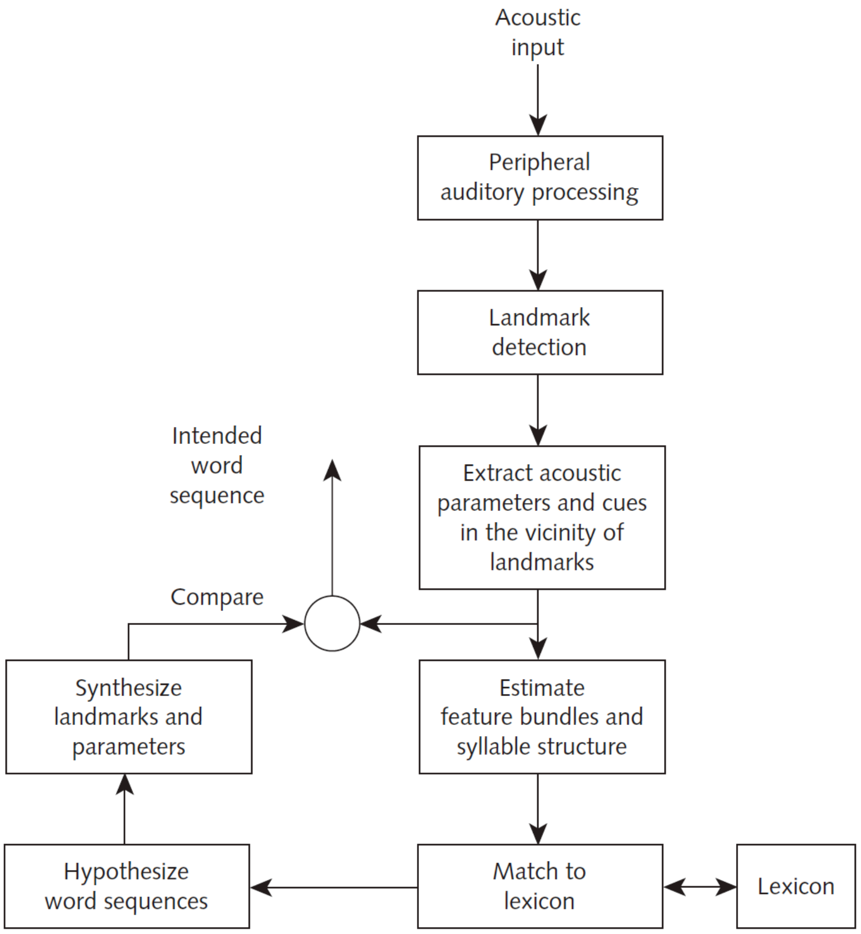

Acoustic input
Peripheral auditory processing
Landmark detection
Extract acoustic parameters and cues in the vicinity of landmarks
Intended word sequence
Compare
Synthesize landmarks and parameters
Estimate feature bundles and syllable structure
Hypothesize word sequences
Match to lexicon
Lexicon

Figure 1.1 Schematic diagram of human lexical access (Stevens, 2008).

Additionally, the bottom-up process is based on the "local analysis of the signal". Given potential variability within the acoustic signal from contextual influences or other sources of variability, some feature bundles might be estimated with a low-confidence level (Stevens, 2008) and results in ambiguous output if no further top-down information is available. The model (2008) also hypothesised that provided sufficient information in syntax and semantics of the sentence in running speech, words could be limited to a small range of choices and if possible be accessed solely based on top-down information without resorting to the precise bottom-up acoustics. The correlation between cloze probability and sentence completion consistency might provide evidence to such hypotheses (Block & Baldwin, 2010; Staub, Grant, Astheimer & Cohen, 2015).

Models of lexical access have traditionally concentrated on segmental processing; how tonal information is processed and how it would fit in a perception model still invites more theoretical and empirical work. Among studies on lexical tone processing, Ye and Connine (1999) investigated the processing asymmetry between segmental and tonal information in a series of vowel–tone detection tasks. The participants were asked to decide whether the heard syllable contained the target combination—/a/ with Tone 2 (the rising tone) in Mandarin Chinese by pressing the button on the keyboard. The non-target stimuli contained either a correct vowel and a wrong tone, i.e. /a/ with Tone 4 (the falling tone), or a correct tone but a wrong vowel, i.e. /i/ with Tone 2. The results revealed significant slower responses to the wrong-tone stimuli than the wrong-vowel stimuli, particularly when they were presented in isolation or with non-predictive context. However, if presented with highly constraining context such as idioms, participants showed no disadvantage in detecting tone mismatch. Their findings clearly suggested contextual influence on tone perception.

To accommodate the context effect on tone processing, Ye and Connine (1999) further proposed modifications of a separate level of toneme to the existing TRACE model. According to their assumptions (Figure 1.2), lexical tone processing operates on a separate level of toneme, whose activation strength is determined by both "the goodness of the input signal and the lexical feedback connections"; and critically, the lexical feedback connection is stronger for tonemes than phonemes (Ye & Connine, 1999). This helped to interpret the results in their study: when the non-target stimulus was embedded as the last syllable in a four-syllable idiom, the stimulus item was highly predictable given the idiomatic context, which sent down strong lexical connection for toneme processing, and the tonal disadvantage was therefore compensated. Moreover, the toneme hypothesis allows partial match or mismatch on the tonal level such that the perceptual system seems tolerant of lexical tone variation in natural speech.

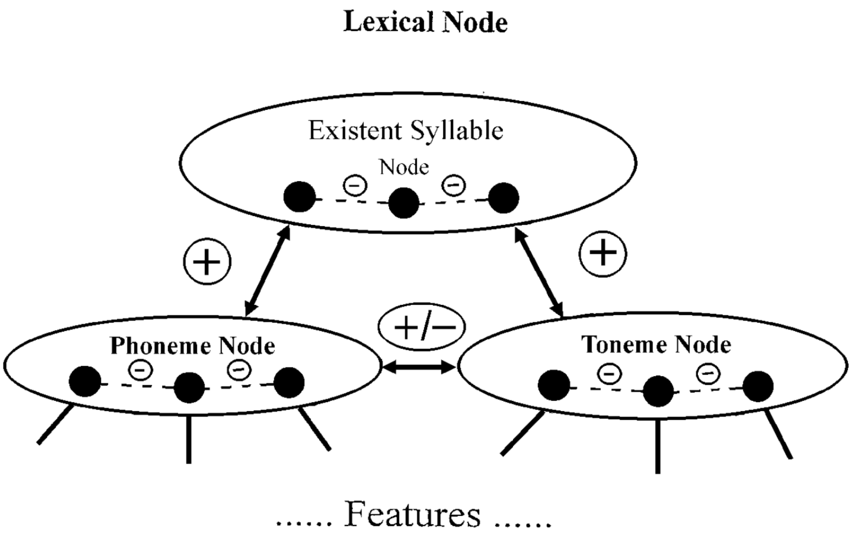


Figure 1.2 Ye & Connine's modified TRACE model from Fig.1 in Ye & Connine (1999), titled as "Phoneme, toneme, and syllable nodes and connections. Larger arrows indicate greater connection strength".

More recently, Gao et al. (2019) tested the tone-to-phoneme disadvantage void of context in a speeded sameness judgement task and found empirical evidence for an additional level of representation, i.e. atonal syllable (Figure 1.3), for lexical access. In the experiment, participants responded to sequentially presented monosyllables with a "same" or "different" response; the stimuli in each trial differed in either the whole word (consonant, vowel and tone), atonal syllable (consonant and vowel), consonant, vowel, tone, or the ear receiving the input. The results confirmed Ye and Connine's (1999) findings concerning the processing disadvantage for tonal information without lexical context. However, in terms of levels of processing, reaction time and error rate data suggested that segmental information can be accessed before integration with tonal information and there might exist an intermediate level of atonal syllable (segments without tones) in lexical access; moreover, both word- and syllable-level information seems readily accessible even before accessing the component phonemes and tones.

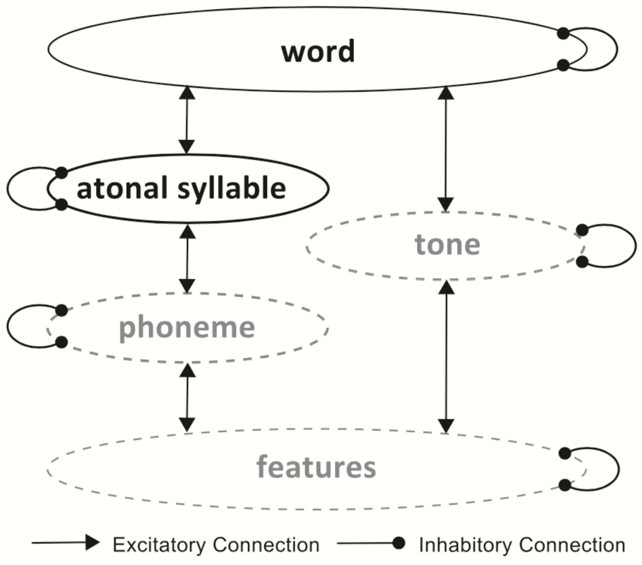


FIGURE 5 | The Reverse Accessing Model (RAM) for sub-lexical phonological processing of tonal speech. Each eclipse indicates a level of processing devoted to a specific type of phonological representation. The solid lines indicate ready access to the extracted information, while the dotted lines mark "hidden" information that can only be accessed by reactivating the system with the perceived word. Information accessing starts from the top and proceeds to the lower level if necessary.

Figure 1.3 The Reverse Accessing Model as FIGURE 5 in Gao et al. (2019).

These findings were incorporated in the Reverse Accessing Model (RAM), where two extra levels of representation were added to the TRACE model (Gao et al., 2019; see Figure 1.3). The level of atonal syllable refers to the combination of consonant(s) and vowel(s) before attaching tonal information; the level of tone represents lexical tone information as from the acoustic signal. The RAM also contrasts between readily available information (word and atonal syllable) indicated by solid lines and accessible-if-necessary information (phoneme and tone) with dotted lines. The model predicts that given the ready accessibility, listeners make better and faster judgments on words and atonal syllables than the component phonemes and tones; phonemes and tones are considered "hidden" in the default state and accessed "if and only if information at the higher level is insufficient for the task at hand" (Gao et al., 2019). Further, they argued for a reverse order of *information accessing* relative to *information* *processing*, unlike the original TRACE model, where information would be accessed as soon as being processing. According to the RAM, information processing follows bottom-up procedures from the acoustic signal, while information accessing starts from the top—listeners first access information of words and then syllables, and that of phonemes and tones is not necessarily accessed unless needed for a specific task.

1.2.2 Discussion and consensus

The models reviewed above demonstrate a few debated perspectives on spoken word recognition: serial search vs. parallel processing, competition vs. non-competition, and interactive vs. non-interactive processing. Comparisons of the models in these aspects are summarised in Table 1.1.

Specifically, the serial search mechanism can be viable for written word recognition since visual information allows instant availability of the word entity as one could read through not only discrete words, but also texts, letters, and choose which part to focus on freely. However, for the spoken word recognition, speech signals are received in a fixed temporal order, which confines the hearer's knowledge of the input to the accumulated amount over time (Taft, 1986). The incremental acoustic information does not necessarily match to a specific lexical item; it seems more likely that multiple candidates are simultaneously processed given the accumulating input. Moreover, the serial search mechanism cannot adequately explain phenomena such as mismatch between input and target word, and frequency effects (Xu & Taft, 2015).

Table 1.1 Summary of the debatable aspects of the reviewed lexical access models. The sign "-" means unspecified or unclear.

	Parallel processing	Competition	Interactive
Autonomous search model	no	no	-
The logogen model	yes	yes	-
The Cohort model	yes	no (original) yes (revised)	-
The activate and check model	yes	yes	-
The TRACE model	yes	yes	yes
The Merge & Shortlist	yes	yes	no
The model of acoustic landmarks and distinctive feature	yes	yes	yes
Ye & Connine's modified TRACE	yes	yes	yes
Reverse Accessing Model (RAM)	yes	yes	yes
The hypothesis of the current thesis	*yes*	*yes*	***yes***

Parallel processing of candidates has been adopted in many models of lexical access in the current literature. Models in support of this assumption include the logogen model (Morton, 1969), the Cohort model (Marslen-Wilson, 1978, 1980), the activate and check model (Taft, 1986), the interactive-activation model (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982), the TRACE model (McClelland & Elman, 1986), Taft's AUSTRUAL model (Taft, 1991, 2006, 2015), the Shortlist (Norris, 1994), Merge (Norris, 1999), the model of acoustic landmarks and distinctive feature (Stevens, 2008), the RAM model (Gao et al., 2019), among others.

Given the parallel processing mechanism, it is likely that activated candidates compete in terms of certain activation level and the one that attains the predetermined threshold is accessed. Three general types of competition can be identified in the previous models: (1) gradual elimination of mismatched candidates given accumulative information, such as the initial version of the Cohort model, (2) frequency-based competition, such as the revised Cohort model and the activate and check model, and (3) feature-based competition, such as the logogen model and the model of acoustic landmarks and distinctive features. The competition mechanisms are more complicated in the connectionist models as they typically allow traces of processed information to flow between connected units across levels of representation.

As for the contrast between interactive and non-interactive approach, a wide variety of models exist that differ on this aspect. Non-interactive models consider lexical processing as entirely bottom-up (e.g. The Merge and Shortlist); interactive models allow bi-directional interactions with top-down feedback from higher levels of representation (e.g. TRACE and its extensions). For tonal speech specifically, empirical studies and current theoretical assumptions strongly suggest top-down lexical influence on tone perception (e.g. Ye & Connine's modified TRACE and the RAM model). Previous research has verified this assumption for native speech; whether it holds for unfamiliar or non-native tonal speech remains unclear. The perception chapters in the thesis will look into the presence of top-down processing for non-native Mandarin speech and expect an affirmative answer. As for the accessibility of tonal information, previous findings have suggested an inferior status of tonal representation particularly without lexical context; the present dissertation will provide further evidence for this topic particularly in Chapter 3.

The following Section 1.3 continues with the topic of interactive processing and highlights the assumption of the constructive nature of speech perception in discussing top-down and bottom-up mechanisms for spoken speech processing. In general, recent research has shown a commonality in favour of integrated mechanisms for speech perception, broadly evidenced or stimulated by empirical studies on general auditory and learning, and perceptual adaptation to non-native and unfamiliar sounds, to be reviewed in Section 1.4.

1.3 Constructive speech perception: integrated top-down and bottom-up processing?

Spoken speech is highly variable in the acoustic information (Section 1.1), but perception is rather accurate and constant (Beddor, 2017). It is rational to speculate that accurate perception of speech sounds may rely on other forms of information in addition to the bottom-up acoustics. It is likely that the perceptual system makes use of all the available information to construct a particular linguistic unit.

Two broad mechanisms for processing spoken speech exist in the current theoretical frameworks: bottom-up processing based on acoustic properties obtained directly from the speech signal and top-down processing using prior linguistic or world knowledge. Certain speech illusions nicely demonstrate the influence of both types of information for speech perception. For example, with the phonemic restoration effect, when listeners hear a word with one phoneme missing and replaced by a brief noise, they are able to mentally restore the absent segment and recognise the word; some may even be unaware of the missing element (Warren & Obusek, 1970). The McGurk effect occurs when listeners receive competing information across audio-visual modality; perception of the target sound depends on the relative weighting between the auditory input and the visual input, which might be affected by the listener's relative experience with the visual and the auditory input. In both cases, the speech perception system draws upon information aside from the acoustic signal in constructing the intended linguist percept when the acoustic signal contains insufficient or ambiguous information.

Previous research has varied views on whether bottom-up and top-down information interacts in a bi-directional approach for lexical access, or they follow a hierarchical structure without interaction. A clear example of an interactive model is the TRACE model of speech perception (McClelland & Elman, 1986). The crucial assumption of the interactive approach is that if bottom-up information is indeed adequate, top-down processing becomes unnecessary and optional; however, when the acoustic signal inadequately matches with a phonological representation, top-down information provides guidance for modification and reconstruction on the output of lower-level processing. Similar assumptions were also proposed in the model of landmark and acoustic features (Stevens, 2008; Section 1.2.1).

A major opponent of the interactive mechanism comes from the studies on the contextual effects on spoken word identification (Connine, 1987; Connine & Clifton, 1987). In Connine's experiment (1987), word pairs manipulating the word-initial consonant (e.g. "tent" vs. "dent") were embedded in either a semantically consistent or a semantically inconsistent sentence; the target words were in sentence-final position. Participants identified the consonant upon hearing the target word; at-boundary reaction times were measured accordingly. They also judged whether the word was consistent with the sentence semantics by the end of each trial. Their evidence for a non-interactive approach was that reaction times did not differ for the ambiguous stimuli between the consistent and inconsistent context; they argued that the effect of sentential context was only "post-perceptual" that happened at a later stage of lexical decision.

The main criticism of the above study is that the consonants being identified in the task were familiar to the participants and present in their native segmental system; therefore, top-down information was primarily used as the predictive context, not for the purpose of retuning the phonological inventory. In this sense, lack of contextual effect cannot entirely obviate integration of top-down process in speech perception. It is possible that as long as participants were confident enough to make the decision based on the acoustics given their native knowledge of the segments, top-down contexts may become less salient in guiding the lexical decision. Further discussion regarding interactive processes in speech perception can be found in the work of McClelland et al.'s (2006) and McQueen, Norris and Cutler (2006).

In fact, instead of fixating on whether it is strictly interactive or non-interactive, insights may arise from research on the relative roles of top-down and bottom-up processing as more recent literature tends to converge on the assumption of integrated mechanisms with both "ascending and descending processes" and the inclusion of general auditory and learning methods in speech perception (Kinchla & Wolfe, 1979; McClelland, Mirman & Holt, 2006; Rauss & Pourtois, 2013).

1.4 Perceptual learning and adaptation

While the perceptual system does a remarkable job processing phonetic variation in native speech, listeners are also capable of perceiving sounds that are novel or unfamiliar provided experience with certain input. How listeners extract useful information from the novel input and whether experience with less-expected sounds exerts influences on the perceptual system are within the scope of research on perceptual learning and perceptual adaptation. The term perceptual learning generally refers to experience- or practice-induced changes in the perceptual system and is often associated with formation of long-term representations, whereas perceptual adaptation targets more immediate and temporary processes (Gibson & Gibson, 1955; Goldstone, 1998; Connolly, 2017; Melguy, 2022). The two terms are sometimes used interchangeably in the current literature, but perceptual learning specifically requires retention of the learning effect after longer periods of time. The present dissertation is more on the side of perceptual adaptation, but literature on both perceptual learning and adaptation is relevant.

It is commonly assumed that perception can be "shaped both by a perceiver's knowledge and by his or her past experience with particular stimuli" (Samuel & Kraljic, 2009). Experience with noncanonical stimuli or variability would presumably lead to perceptual changes, be it long-lasting or provisional. What are changed and in which form, how long these changes may last, and which type of and how much amount of input would induce such changes have been the key questions researchers endeavour to unpuzzle. Particularly, to evaluate the changes in the perceptual system, two major approaches were summarised in Samuel and Kraljic's (2009) review. One focuses on the measurable improvement in perceiving unfamiliar speech stimuli. The other one is concerned with how perceptual learning/adaptation reshapes the phonetic space and changes the way that contrastive sounds are categorised. The following Sections 1.4.1 and 1.4.2 introduce the two approaches respectively with relevant studies.

1.4.1 Perceptual learning/adaptation as overall improved perception

Early studies see perceptual learning as the improved ability to identify or discriminate stimuli in unfamiliar speech (Gibson & Gibson, 1955; Goldstone, 1998). Considerable improvement after training was entailed in the previous definitions of perceptual learning. Gibson & Gibson (1955) described the learning effect as being “more sensitive to the variables of the stimulus array”. Goldstone (1998) characterised perceptual learning as improvement in listeners’ responses to the readily environment and suggested four general mechanisms for perceptual learning—“stimulus-detecting” for recognition of the unfamiliar stimuli, “differentiation” for identification of the critical differing cues between stimuli, “unitisation” for inclusion of multiple cues for better perception, and “attention-weighting” for assigning relative weight to particular stimuli based on experience and exposure.

Previous studies assuming this improvable aspect have investigated both the perception of novel phonetic contrasts and the perception of utterances in unfamiliar speech. Specifically, Lively, Logan and Pisoni (1993, 1994) trained native Japanese speakers to discriminate a non-native phonemic contrast between /ɹ/ and /l/ in English in a three-week session. The listeners were tested in either a high-variability (multiple-talker) or a low-variability (single-talker) condition. The results revealed significant overall improvement in the sound discrimination task. With single-talker stimuli, listeners generalised the learned contrast to new words, but not to new talkers; with multiple-talker stimuli, listeners generalised to both new words and new talkers. Moreover, persistent learning effect was found after six months.

For the adaptation to accented and dialectal speech, Bradlow and Bent (2008) also investigated the effect of input variability on native English listeners’ perception of Chinese-accented speech in the sentence transcription task and found similar patterns as in Lively et al.'s (1993, 1994) study—exposure to one speaker's speech confined generalisation to that particular speaker. But crucially, they found that listeners were able to generalise to novel speakers if they received increased amount of exposure from a particular speaker, which may compensate for the loss from one speaker.

In another study, Flege (1995) tested Mandarin speakers' discrimination of syllable-final /t/ and /d/ in English after two weeks' training in two different tasks—a "yes/no" discrimination task and a sameness matching task. The results suggested no between-task differences and considerable improvement in accuracy in both tasks. Apart from learning of segmental contrast, perceptual learning/adaptation also occurs in the suprasegmental system. In Wang, Spence, Jongman and Sereno's (1999) study, native American English speakers were trained to identify the four lexical tone categories of Standard Mandarin using naturally produced words by native Mandarin speakers over eight training sessions in two weeks. Listeners improved by around 20% in accuracy after training; the improvement also generalised to new speakers and new stimuli, and was retained after six months (Wang et al., 1999).

The learning effect in the above studies was observed after a relatively long-period of training. However, Clarke and Garrett (2004)'s study showed that adaptation to foreign-accented speech occurred after a brief period of exposure about one minute. In their study, four blocks of Spanish- and Chinese-accented sentences were used as the stimuli; particularly, the sentences were of low probability for the sentence-final word. Listeners were presented with the spoken sentences with the final word in each sentence probed by either a matching or a mismatching visual object; error percentage and reaction time were measured for listeners' "yes/no" responses to whether the heard final-words matched the visual objects. The major finding of the study was that listeners adapted to the accented speech after exposure to the first three blocks which lasted about one minute; in some conditions, adaptation even occurred after two to four sentences' exposure.

1.4.2 Perceptual learning/adaptation as reshaped phonetic space

While overall improved perception clearly indicates the presence of learning effects and adpatation to unfamiliar speech, questions still remain on whether such effects reflect actual changes in the phonetic space, or they merely result from general auditory and learning processing of the distributional information in the input. If there are indeed perceptual changes, how would the reshaped phonetic space look like? To answers questions like this, the other approach for measuring the exposure-induced perceptual change focuses on the topic of phonetic retuning, also termed as phonetic recalibration (Norris, McQueen & Culter, 2003; Bertelson, Vroomen & De Gelder, 2003; Samuel & Kraljic, 2009).

The general hypothesis is that listeners tend to retune their phonetic space to better accommodate the received input. The often-implemented procedure is to present listeners with ambiguous sounds predictable by certain lexical context and then test their identification of the ambiguous sounds in a sound continuum to see whether their judgment is biased towards the context-predicted end (Norris, McQueen & Cutler, 2003; McQueen, Norris & Cutler, 2006; Clark-Davidson, Luce & Sawusch, 2008; Samuel & Kraljic, 2009).

Specifically, in Norris et al.'s (2003) study, Dutch listeners received input of an ambiguous sound between [s] and [f] with highly predictive lexical context and were then tested to categorise a continuum of [s]-[f] sounds into two categories. The ambiguous sound was inserted in the word-final position to replace either [s] or [f] in real Dutch words, so that the manipulated stimuli containing a final ambiguous sound were lexically biased towards either a [s] recognition or a [f] recognition. In the initial lexical decision task, listeners heard either normal final-[s] words and ambiguous final-[f] words, or normal final-[f] words and ambiguous final-[s] words, and judged whether the heard word was a real word in Dutch. In the following categorisation task, listeners were presented with a continuum of [s]-[f] sounds and reported each sound as either [s] or [f]. The results revealed a strong lexical influence on sound categorisation. Specifically, lexically biased input led to a boundary shift in the phonetic space: listeners identified more intermediate sounds in the continuum as the one which was ambiguous, but lexically predicted in the preceding input.

Since a lexical decision task was used to present the training stimuli in Norris et al.'s (2003) study, recognition of the word identity was inevitably involved during exposure. This design poses a question: is it the explicit lexical decision that instigates the learning process? Does the change in the phonetic space necessarily require accurate lexical decision? In fact, the results suggested otherwise—the listeners identified most of the ambiguous words as real words instead of expected non-word responses. To answer the former question, a follow-up study by McQueen, Cutler and Norris (2006) replicated the basic design of the previous study, except that listeners were asked to count the number of the trials including [s] or [f] without attending to the lexical identity of the word in the training session. Their results showed similar findings as Norris et al.'s (2003). Both studies indicated that though sound categorisation can be guided by lexical information, the change in the phonetic space relies on neither the accurate lexical decision nor the mere presence of lexical processing, and perceptual learning of novel phonetic contrast is essentially automatic.

A similar design was implemented by Kraljic and Samuel (2005) and Eisner and McQueen (2006). The two studies further investigated whether the categorisation shift between [s] and [f] remained for longer periods. The results showed that the learning effect was present in twenty-five minutes after the initial experiment (Kraljic & Samuel, 2005), and still existed even after twelve hours (Eisner & McQueen, 2006).

In a more recent study on phonetic recalibration, Melguy (2022) trained native English speakers with spoken words containing an ambiguous sound [θ]-like sound between [θ]-[s] and tested them to categorise sounds in either a [θ]-[s] continuum or a [θ]-[f] continuum. The purpose of using different test continuums was to examine whether the phonetic recalibration is caused by category shift as assumed in many previous studies, or it is due "uniform broadening" of the ambiguous category (Melguy, 2022). Since the input was lexically biased towards the ambiguous [θ] relative to an unambiguous [s] sound, if the mechanism is by category shift, listeners would categorise more sounds as [θ] in the [θ]-[s] continuum, but show no difference in the [θ]-[f] continuum as this contrast received no lexical influence and should remain intact; however, if it is general broadening of the phonetic space for [θ]-like sounds, listeners would report more [θ] sounds in both [θ]-[s] continuum and [θ]-[f] continuum. The results conformed to the former expectation of category shift, which suggested actual changes in the phonetic space after exposure to noncanonical phonetic contrast.

A minor drawback of the above studies is that it is unclear whether the observed adaptation (i.e. reshaped phonetic space) would necessarily lead to improvement in lexical decision. While listeners readily adjusted categorisation boundaries after exposure, they were not tested on whether their judgment of words and nonwords improved over the training trials or after the exposure. Maye, Aslin and Tanenhaus (2008) otherwise examined both boundary shift and post-exposure lexical decision in the study on English speakers' adaptation to an experimentally manipulated English accent. In Maye et al.'s (2008) study, listeners received passage exposure of the story "Wizard of Oz" in standard American English accent and then in a modified accent with lowered front vowels (e.g. [i] pronounced as [ɪ], [ɪ] pronounced as [ɛ]). After each exposure session, listeners were tested in an auditory lexical decision task using normal-accented words for the expected word responses and vowel-lowered words for non-word responses. The results showed that listeners successfully adapted to the modified accent with a larger proportion of the modified words being identified as legitimate after hearing the modified passage. Crucially, listeners generalised the heard lowering pattern from front vowels to back vowels as well. It seemed that listeners tend to internalise the newly adapted feature to modify a larger natural class of speech sounds (Maye et al., 2008).

1.4.3 Potential factors in perceptual learning/adaptation

Previous studies have generally agreed on listener's ability to construct the perceptual system to become more attuned to the ambient environment (Munro & Derwing, 1995; Weil, 2001; Norris, McQueen, & Cutler, 2003; Clarke & Garrett, 2004; Zheng et al., 2005; Bradlow & Bent, 2008). This section summarises a few notions that have been tested or hypothesised concerning potential factors in perceptual learning/adaptation.

To begin with, perceptual learning/adaptation seems input specific. Listeners do not randomly adapt to any novel input whenever it becomes available; instead, they are more sensitive to particular features in the stimuli that are easier to map onto the existing distinctions in the sound system (see Watanabe, Nanez & Sasaki, 2001; Seitz et al., 2005). According to Kraljic, Brennan and Samuel (2008), features that listeners choose to adapt to are those supplementary to the existing system of the phonological features, rather than newly created distinctions. Moreover, how well listeners adapt to unfamiliar phonetic contrasts seem dependent on their distinctiveness relative to the contrasts in listeners' native speech (Kraljic et al., 2008). Similar assumptions can be found in certain hypotheses for second language acquisition, such as Best’s (1994) Perceptual Assimilation Model and Flege’s (1995) Speech Learning Model (Samuel & Kraljic, 2009).

Adaptation outcome may also be modulated by talker-variability in the exposure stimuli. Differing results have been found in adaptation with single-talker exposure and multiple-talker exposure (Clarke & Garrett, 2004; Floccia, Goslin, Girard & Konopczynski, 2006; Bradlow & Bent, 2008). Low talker-variability exposure may help direct listeners’ attention to the differing patterns in speech itself, rather than talker-specific characteristics (Floccia et al., 2006). High talker-variability exposure may provide more information which potentially allows listeners to normalise across-talker differences and generalise to new talkers (Clarke & Garrett, 2004). Previous findings also suggested that generalisation to new speakers is also possible with input from just one speaker, as long as there is sufficient amount of exposure from that speaker (Bradlow & Bent, 2008).

In fact, results on the amount of exposure sufficient to induce perceptual learning/adaptation vary in different studies. Bradlow & Bent's (2008) study showed that listeners’ perception of the accented-speech significantly improved after twenty-one trials of sentence exposure. Reaction time results in Norris et al.’s (2003) study indicated faster responses in lexical discrimination with increased amount of input (Norris et al., 2003). A working hypothesis is that exposure which provides an adequate sampling information of the features would accelerate perceptual processing (Bradlow & Bent, 2008). However, other studies also demonstrated relatively rapid adaptation after minutes’ exposure (e.g. Clarke & Garrett, 2004). A more extreme case with minimal amount of acoustic exposure can be found in Wiener and Ito’s (2016) study. Their study was not particularly on perceptual adaptation, but it is reviewed here as it experimented on the required amount of acoustic information for word and tone identification.

In spoken word recognition, there might exist a period of time from the speech onset, during which the acoustic input is still too vague to decipher. Wiener & Ito (2016) investigated whether impoverished tone acoustics would still trigger accurate tone processing. Twenty-four unique CV(X) syllables contrasting in relative syllable frequency were used as the auditory stimuli; half of them were high-frequency syllables and half were low-frequency syllables based on token frequency in SUBTLEX-CH (Cai & Brysbaert, 2010). For each syllable item, two words were constructed matching to either a most probable tone category or a least probable tone category given tone token frequency in SUBTLEX-CH. For example, for the syllable "shi", the word "shi4" occurs about 50% of the time in the corpora, while "shi3" occurs 6% of the time (Wiener & Ito, 2016). A total of 48 words (24 syllables × 2 tone-probability conditions) were further modified using a gating paradigm. For each word, the sonorant portion of the syllable was divided into 8 gates, each with a 40-ms increment. In the experiment, listeners first heard all the words with the 1-gate increment, and then all the words with 2-gate increments, until all the words of 8-gate increments. In each trial, listeners reported the syllable and the tone category in Pinyin (Wiener & Ito, 2016).

According to Wiener & Ito (2016), in this syllable-frequency by tone-probability design, it was expected that high syllable-frequency might delay listeners' prompt identification of the tone category, while for less frequent syllables, listeners would identify the tone category more promptly, but their identification might be biased towards a more probable tone category upon initial input. However, the results suggested that both lexical frequency and tonal probability had limited impact on tone identification. In general, listeners became more accurate in both syllable and tone identification over the incremental gates across all the conditions. Crucially, they were able to make use of impoverished acoustic information as early as the first 40 ms into the sonorant part to identify the tone category with significant higher accuracy than syllable and word identification. It was therefore hypothesised that suprasegmental processing may happen in parallel to segmental processing; prediction of the tone category is made upon the initial input of the segmental information.

This dissertation will focus on a subset of the mentioned factors. Particularly, Chapters 3 and 4 investigate the quality and quantity of exposure in adaptation to the unfamiliar tone systems, as well as the effect of tonal distinctiveness relative to listeners’ native tones.

1.5 The terms in Standard Mandarin and Mandarin dialects

As the dissertation focuses on Mandarin Chinese dialects, it should be helpful to establish a few terms and transcription conventions that will be used throughout. Phonologically, all dialects of Mandarin Chinese make use of both segmental (vowels and consonants) and supra-segmental (tone) features to contrast meaning. The standard orthographic system used by the community is character-based, e.g. 啊 (ā); however, the language is commonly transcribed using Pinyin, which uses the Roman alphabet to represent segmental information and a special set of tone markers on vowels to represent lexical tone categories, e.g. ā (Tone 1), á (Tone 2), ǎ (Tone 3), à (Tone 4). To make it easier for readers of non-tonal languages, studies on Mandarin tones have alternatively used numbers to replace the Pinyin tone markers[2]. For instance, the monosyllabic word 花, meaning “flower(s)” in English, can be transcribed as *huā* or *hua1* in Pinyin; the number “*1*” refers to Tone 1. The present dissertation will take the latter approach for specifying tone category and predominantly make use of Pinyin transcription to represent the shared phonological representations across dialects. The exception will be in Chapter 2 on the acoustic-phonetic realisation of Mandarin dialects,

[2] See different transcription systems for Mandarin lexical tones in *The Phonology of Standard Chinese* (Duanmu, 2007, pp. 225-228).

where the IPA and Chao tone numerals will be used to better capture the phonetic differences of the dialects.

1.6 Thesis outline and research questions

To understand how the dialectal speech of Mandarin is perceived and how listeners deal with variability in the lexical tone system, the present dissertation addresses three broad questions in the research chapters:

1. To what extent do the lexical tones differ in their phonetic realisation across Mandarin dialects?

2. What are the perceptual mechanisms for processing the phonetic tone variation?

3. What are the potential factors affecting perceptual adaptation to an unfamiliar tone system?

For the layout of the thesis, Chapter 1 introduces the relevant background literature on the lack-of-invariance problem, the existing models for lexical access, and perceptual learning and adaptation, as well as previous research on the perception of dialectal and tonal speech. The higher-level theoretical assumptions were reviewed in the chapter; more specific reviews can be found in the introduction sections in the following research chapters.

Chapter 2 focuses on the production of lexical tones in the six representative Mandarin dialects—Beijing, Chengdu, Jinan, Taiyuan, Wuhan, Xi'an Mandarin, which demonstrates considerable phonetic variability in the lexical tone systems of Mandarin dialects. The purpose of this chapter is to provide acoustic-phonetic descriptions of the dialect-specific tone categories through corpus-phonetic analysis and statistical modelling. This would tentatively update current knowledge of the tone inventories of the six dialects. The results for dialectal differences of the tone systems could further inform patterns found in the following perception experiments.

Chapter 3 investigates native Standard Mandarin speakers' perception of the familiar (Standard Mandarin) and unfamiliar (Chengdu Mandarin) tone systems with incidental exposure in the sentence semantic plausibility judgment task. The results strongly indicated listeners' rapid adaptation to the unfamiliar Chengdu tones with about two-minute incidental exposure directly from the experimental trials. They were able to process novel tone acoustics even through the lexical decision was made primarily based on the top-down sentential context. Moreover, presentation of the contrastive tone categories in the stimuli was not necessary to induce adaptation to the unfamiliar tone system.

Chapter 4 follows up on the findings in Chapter 3 and examines the effect of explicit passage exposure on adaptation to unfamiliar Chengdu Mandarin and Jinan Mandarin tone systems. We found that perception improvement with explicit exposure was modulated by the relative similarity of the tone system to the listeners' native tone system; more dissimilar tones are easier to adapt to.

Chapter 5 summarises the major findings of the thesis and provides further discussion relative to the theoretical framework for speech perception. Limitations of the present studies and future directions are included in the conclusion.

Chapter 2 Production of Mandarin dialects: comparable segments and distinct tones

2.1 Introduction

Mandarin regional dialects vary considerably in their phonetic realisation, especially in the tone system; however, little empirical research has been conducted to validate these impressions. Chapter 2 focuses on the acoustic phonetics of six Mandarin dialects—Beijing, Chengdu, Jinan, Taiyuan, Wuhan and Xi'an Mandarin, aiming to empirically investigate the comparability and distinctiveness in the vowels and lexical tones of Mandarin dialects. Speech data was collected from each dialect using remote smartphone data collection techniques; each vowel space and tonal inventory was then analysed and compared using corpus-phonetic methods. The present production study also provides a practical pipeline of corpus construction with remotely collected speech data, as well as annotations for future research on Mandarin dialect variation.

Before diving into the investigation, the introduction will present an overview to the relevant foundational topics for understanding and analysing phonetic variation across Mandarin dialects. Section 2.1.1 reviews previous classifications of Chinese dialects and introduces the Mandarin language group and Mandarin dialects that differ primarily in the phonetic domain. Sections 2.1.2 and 2.1.3 further elaborate on the sound systems of Standard Mandarin and Mandarin dialects in terms of the known segmental system (Sections 2.1.2) and lexical tone inventories (Sections 2.1.3). Section 2.1.4 clarifies the acoustic parameters of lexical tones to be measured and analysed in the production study in this chapter. Section 2.1.5 reviews the previous attempts to quantify the relative similarity of the sound systems between dialects. Additionally, Section 2.1.6 reviews speech data collection methods and current speech corpora related to the Mandarin language, which serves as a foundation for the new methods involving remote audio data collection that had to be implemented during the pandemic. The rest of the chapter follows the structure of a typical empirical paper. The method section describes the participants, recording materials, procedure of the production experiment, and the detail on corpus construction and acoustic analysis. The result section presents the analyses of the vowel spaces and the tone inventories of the six dialects. The chapter ends with a summary of the major findings, followed by the statement of the authorship and publication status.

2.1.1 Mandarin dialects as a unique dialect group

The Mandarin language is one of the language groups spoken by around 70% of the population in mainland China and is rich with regional and other sociolinguistic variation. Mandarin dialects broadly refer to the regional and local varieties of Mandarin. They were primarily associated with the Han ethnic group, the largest ethnic group in China, and have been widely recognised as an extensive dialect group in the current literature since the first scientific publication on Chinese dialect classification by Li Fang-Kuei (1937, 1973). Specifically, the handbook *Xiandai Hanyu Fangyan Gailun* (《现代汉语方言概论》, "*The Modern Outline of Chinese Dialects"*, Hou, 2002) specified nine major groups of Chinese dialects—*Mandarin*, *Wu, Xiang, Gan, Yue, Min, Hakka, Jin* and *Hui* in. In the *Language Atlas of Chinese Dialects* (《汉语方言地图集》, Wurm, Li, Baumann, & Lee, 1987; Figure 2.1), an additional *Ping* group was mapped alongside the above nine groups, but the *Ping* language is often considered integral to the *Yue* language. Moreover, a seven-group scheme was proposed in *Hanyu Fangyan Gaiyao* (《汉语方言概要》, "*An Outline of the Chinese Dialects"*, Yuan, 1960), where *Hui* was considered a transitional group between *Mandarin* and *Wu*, and *Jin* was incorporated into the *Mandarin* group. Though debates remain concerning dialect typology, Mandarin dialects (also termed as Mandarin-group or Mandarin-branch dialects) have been studied and analysed as a unique near-homogenous group in sociolinguistic studies since the 1940s (Chao, 1943; Wurm et al., 1987; Hou, 2002; Ho, 2003; Szeto, Ansaldo & Matthews, 2018). More references can be found in others handbooks and reports, such as *Languages and dialects in China* (Zhao, 1943), and dictionaries of Chinese dialects, such as *Hanyu Fangyin Zihui* (《汉语方音字汇》, "*The Sounds of Chinese Dialects*", 1989) and *Xiandai Hanyu Fangyan Dacidia* (《现代汉语方言大词典》, "*The Modern Dictionary of Chinese Dialects*", Li, 2002).

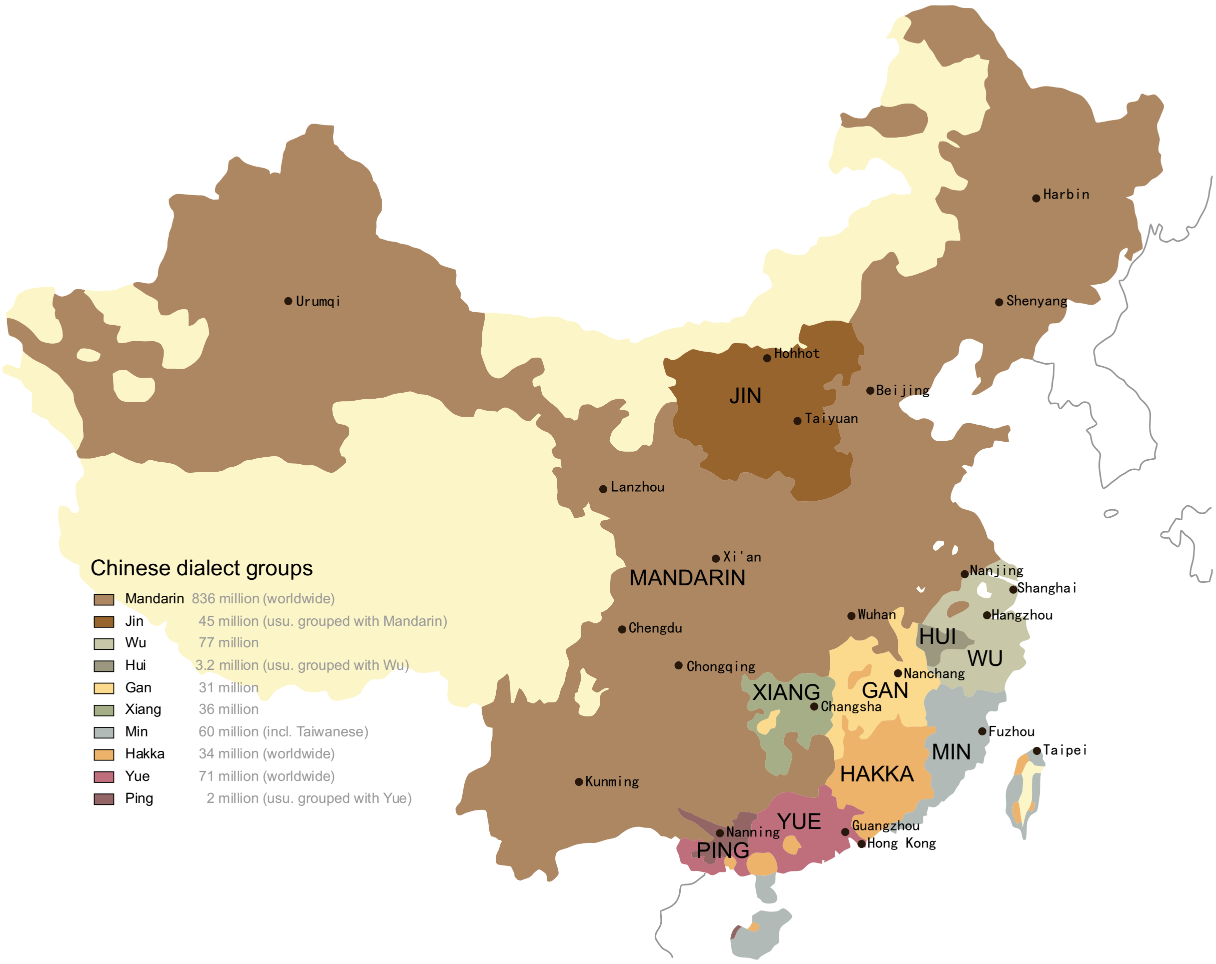


Figure 2.1 Map of Chinese dialect groups based on *Language Atlas of Chinese Dialects* (Wurm et al., 1987) by Wyunhe (2011). The Mandarin group is in light brown.

Within the Mandarin group, Mandarin regional dialects cover a vast geographic area compared to the non-Mandarin groups and contain several subgroups (Figure 2.1). Early classification specified eight subgroups of Mandarin dialects: Beijing, North-eastern, Jiaoliao, Jilu, Zhongyuan, Lanyin, South-western and Jianghuai regional Mandarin (Wurm et al., 1987; Hou, 2002). Considering fusion of ethnic groups and language contact, current literature tends to group Mandarin dialects more broadly into fewer sub-areas given their geographic vicinity—the four general sub-areas include Northern Mandarin (e.g., Beijing, Jinan, Jilin Mandarin), North-western Mandarin (e.g., Taiyuan, Xi'an, Lanzhou Mandarin), South-western Mandarin (e.g., Chengdu, Wuhan, Kunming Mandarin), and Eastern Mandarin spoken in Anhui province and part of Nanjing (Norman, 2013).

The production study in this chapter selected two mainstream dialects from each of the first three sub-areas: Beijing, Chengdu, Jinan, Taiyuan, Wuhan and Xi'an Mandarin. Eastern Mandarin dialects were not included as they are heavily influenced by *Wu* dialects, a non-Mandarin language group. Beijing Mandarin is a regional variety of Mandarin spoken in Beijing and its near suburbs, and marginally different from Standard Mandarin[3], the official language for education and administration in mainland China (see more in Sections 2.1.2 and 2.1.3). For Taiyuan Mandarin specifically, whether it is a Mandarin dialect or belongs to the *Jin* group is still controversial and both versions exist in the literature. Figure 2.1 indicates that Taiyuan Mandarin belongs to the *Jin* group which chiefly differs from the Mandarin group in the presence of checked tones—Mandarin dialects typically do not have checked tones in their sound inventories, whereas other sources consider Taiyuan as a tentative member of Mandarin dialects given their linguistic similarities in general (Yuan, 1960; Norman, 2003; Tang & van

[3] Though Beijing Mandarin provides the basis for the phonology of Standard Mandarin, Gui and Liu (2011) summarised nine intrinsic differences between Beijing Mandarin and Standard Mandarin, such as the particular set of vocabulary, rhoticity, vowel nasalisation, as well as speech rhythm. See also in Chirkova & Chen's (2011) work on Beijing Mandarin.

Heuven, 2009). Intelligibility studies also suggested that Mandarin speakers understood Taiyuan speech with over 70% of accuracy in the word translation task, similar to the understanding of Mandarin dialects in general and considerably higher than the non-Mandarin groups (Tang & van Heuven, 2009). The current dissertation therefore left the status of Taiyuan Mandarin unresolved and included it as a representative regional dialect for the North-western subgroup given its high intelligibility for Mandarin listeners.

Previous descriptions of the linguistic characteristics of the Mandarin language have identified several shared features of Mandarin dialects on multiple linguistic levels, including phonetics, phonology, morphology and syntax (Yuan 1960; Hou, 2002; Li, 2002; Norman, 2013). These features would suffice as the criteria for language classification, differentiating *Mandarin* from the other language groups such as *Wu* and *Yue*, but current literature is lacking in a comprehensive evaluation of the linguistic similarity and dissimilarity of Mandarin dialects, especially for their sound systems. Previous studies generally agree that although they share considerable similarity in the phonology, lexicon and syntax, Mandarin dialects primarily differ in their phonetic inventories (Chao, 1943; Ho, 2003; Szeto et al., 2018), especially in the lexical tone inventories (Wu, Chen, van Heuven, & Schiller, 2016; Li, Best, Tyler, & Burnham, 2020). Li (2002) and Hou (2002) categorised Mandarin dialects under the notion that the intrinsic difference between Mandarin dialects lies in their phonetic inventories, not in the phonology.

2.1.2 Segmental systems: Standard Mandarin and Mandarin dialects

2.1.2.1 Standard Mandarin segmental system

Standard Mandarin (also known as *Putonghua* or *Guanhua* in Pinyin) is the official dialect promoted nationwide as the lingua franca in China. It was developed based on the Northern Mandarin dialects and the sound system of Beijing Mandarin. The segmental inventory of Standard Mandarin has been well-examined in the earlier literature (e.g. Howie, 1976; Duanmu, 2007; Lin, 2014). Table 2.1 presents the consonant inventory of Standard Mandarin (Yuan, 1960; Duanmu, 2007; Norman, 2013), also considered as the maximal consonant inventory of Mandarin dialects (Norman, 2013). All the obstruent sounds are voiceless; the plosives and affricates of the same place of articulation contrast in aspiration. The maximally complex syllable structure of Standard Mandarin is (CG)V(V) or (CG)V(N) (Duanmu, 2007; Lin, 2014); G is for glides, C for consonants, V for vowels and N for nasals; the parentheses denote optionality. The syllable onsets make use of all the consonants except /ŋ/; the codas are limited to the non-labial nasals. Table 2.2 provides the corresponding Pinyin symbols of the consonants.

Table 2.1 The consonant inventory of Standard Mandarin (and Mandarin dialects).

	Labial	Alveolar	Retroflex	Palatal	Velar
Plosives	p p^h	t t^h			k k^h
Affricates		ts ts^h	tʂ tʂ^h	ç tç^h	
Nasals	m	n			ŋ
Fricatives	f	s	ʂ	ç	x
Other sonorants	(w)	l	ɹ	(j)	

Table 2.2 Pinyin representation of the consonant inventory of Standard Mandarin (and Mandarin dialects).

	Labial	Alveolar	Retroflex	Palatal	Velar
Plosives	*b* *p*	*d* *t*			*g* *k*
Affricates		*z* *c*	*zh* *ch*	*j* *q*	
Nasals	*m*	*n*			*ng*
Fricatives	*f*	*s*	*sh*	*x*	*h*
Other sonorants	*(w)*	*l*	*r*		

For the vowel system, longstanding controversy exists concerning the underlying vowel inventory of the Mandarin language; critically, the number of proposed vowel phonemes varied in different studies (Zhou, 1999; Lin, 2014). The minimal contrastive vowel system contains five monophthongs: three high vowels, one mid vowel (also represented as an unspecified vowel /E/ in other studies) and one low vowel, i.e. /i, y, u, ə, a/ (Duanmu, 2007; Lin, 2014; Table 2.3). Variant surface realisations exist especially for the mid and low vowels. According to Lin (2014), the mid vowel /ə/ has allophonic realisations such as [e, o, ə, ɤ] and the low vowel /a/ can be phonetically realised as [a, ɑ, æ, ɛ]. Lin (2002) also proposed mid-vowel assimilation rules to further explain the common variation of the Mandarin mid vowel. In Table

2.3, the mid back vowel /o/ is placed within parentheses as it is explicitly used in the Chinese Pinyin system, while its phonological status stays unresolved. For the high vowels, the allophones to /i/ are known as apical vowels which occur directly following coronal and retroflex sibilants[4], i.e. / s, ʂ, ts, ts^h, tʂ, tʂ^h/. Standard Mandarin also has a rich cohort of diphthongs and triphthongs (e.g. /ia/, /ua/, /iao/, /uei/), and many of them can be followed by a nasal coda (e.g. /ian/, /uan/).

Table 2.3 The vowel inventory of Standard Mandarin (left) and the Pinyin representation (right).

	Front	Central	Back
High	i y		u
Mid		ə	(o)
Low		a	

	Front	Central	Back
High	*i ü*		*u*
Mid		*e*	*(o)*
Low		*a*	

2.1.2.2 Current state of knowledge on dialect segmental inventories

Though yet under investigation, previous studies have generally agreed that the segmental systems of Mandarin dialects bear a strong resemblance to that of Standard Mandarin (Ho, 2003; Norman, 2013; Szeto et al., 2018). A strong and commonly made assumption is that Mandarin dialects can be analysed assuming the *same* segmental inventory

[4] There were varied views on whether syllables such as /si/ in Standard Mandarin contain an apical vowel or an apical consonant [ʂ̩]. Empirical findings on the articulatory movements of zhi and zi (Chen et al., 2015) suggested that zhi contained a retroflex consonant and a true apical vowel, whereas in zi the vowel was the voiced extension of the apical consonant.

with Standard Mandarin, and meanwhile demonstrate phonetic variation across dialectal varieties (Ho, 2003; Norman, 2013). This assumption has not been fully established and is part of what the current chapter intends to investigate; further research and large amount of dialectal speech data are, however, essential to make such generalisations.

Recent years have seen more acoustic-phonetic studies on the sound systems of Mandarin dialects. The vowel system of Beijing Mandarin typically involves a rhotic version of the mid vowel /ɚ/, which is more extensively used than in Standard Mandarin (Chirkova & Chen, 2011). For Chengdu Mandarin, Huang and Gu (2014) collected recordings of monosyllabic words from ten native speakers of Chengdu Mandarin and plotted the F1–F2 vowel spaces contrasting in three age groups. Hu and Zhang (2018) investigated a particular phenomenon of vowel raising in Chengdu Mandarin using production of words containing /an/ by seventeen native speakers (see also in Li & Hu, 2023). They found a general raising effect of /a/ towards an [ɛ] sound among the younger speakers, while the elder speakers were more likely to produce the low vowel as [æ]. They also took F1 and F2 measurements from the production of monosyllables with a fixed onset of /t/ to demonstrate talker-specific vowel spaces. For Jinan Mandarin, Yang (2011) conduced an acoustic-phonetic analysis on the vowels of Jinan Mandarin and observed a triangular acoustic space with /i, a, u/. A similar study of vowel acoustics was done on Taiyuan Mandarin (Xia & Hu, 2016), which observed the vowel space of the monophthongs and the formant dynamics of the diphthongs. Zhou (1999) specified five contrastive vowels /i, E, ə, a, o/ for Wuhan Mandarin; /E/ was considered an unspecified front vowel. Li and Wu (2016) investigated the frequencies and distributions of the initials and finals in Xi’an Mandarin. Nonetheless, current literature still lacks acoustic-phonetic analyses on Wuhan and Xi’an segmentals. Segmental analyses of other Mandarin dialects, such as Tianjin, Zhushan and Hefei Mandarin, were published as IPA illustrations (Li, Chen & Xiong, 2019; Chen & Guo, 2022; Kong, Wu & Li, 2022). The specific vowel figures can be viewed in the relevant papers; a shared feature of the Mandarin vowel spaces is that they unitarily exhibit a triangular shape with the boundary points marked by two high vowels and a low vowel.

Though many studies have looked into Mandarin dialect sound inventories individually, not much investigation has been done in the comparative manner to verify uniformity of the segmental system of Mandarin dialects. Moreover, inconsistent symbols used for the phonological representations in different studies have posed difficulty in making direct comparisons and generalisations. For example, Zee and Lee (2007) included /ɤ/ in an inventory of /i, y, ə, ɤ, a, u/ for Standard Mandarin. But [ɤ] was considered an allophone of /ə/ in Duanmu (2007)'s representations, where [o, e, ə, ɤ] were all allophonic variations of the schwa /ə/.

A plausible solution of mapping out comparable sound inventories across dialects is to approximate a maximal overall pattern with a limited number of anchor segments, which could be applied to most if not all Mandarin dialects (Duanmu, 2007). Particularly for the vowel inventories, the "vowel-less" approach (Lin, 2014) was proposed from the perspective of nonlinear phonology, by which multiple allophones are made available through spreading or assimilating features. For the case of Mandarin dialects, a weaker version as to the phonological inventory of Mandarin dialects is adopted in this study—Mandarin dialects have *comparable* segmental inventories. For the vowel inventory specifically, /i, ə, a, u/ were used as anchor vowels to represent the periphery and the centre of the F1–F2 acoustic space for vowel analysis across the dialects.

2.1.3 Lexical tone inventories: Standard Mandarin and Mandarin dialects

2.1.3.1 Standard Mandarin tone system

Regarding the tone system, Standard Mandarin has four lexical tone categories and an additional neutral tone. The textbook description of the four-tone system in Standard Mandarin is shown in Table 2.4. Standard Mandarin Tone 1 is marked in Chao tone numerals as [55], Tone 2 as [35], Tone 3 as [214], and Tone 4 as [51]. Figure 2.2 illustrates the schematic tone contours of Standard Mandarin based on Chao tone numerals.

Table 2.4 The tone system of Standard Mandarin: examples with the syllable /ma/.

Character	Pinyin	Tone category	Chao tone numerals	Gloss	Pinyin transcription
妈	mā	Tone 1	[55]	“mom”	*ma1*
麻	má	Tone 2	[35]	“linen”	*ma2*
马	mǎ	Tone 3	[214]	“horse”	*ma3*
骂	mà	Tone 4	[51]	“to scold”	*ma4*

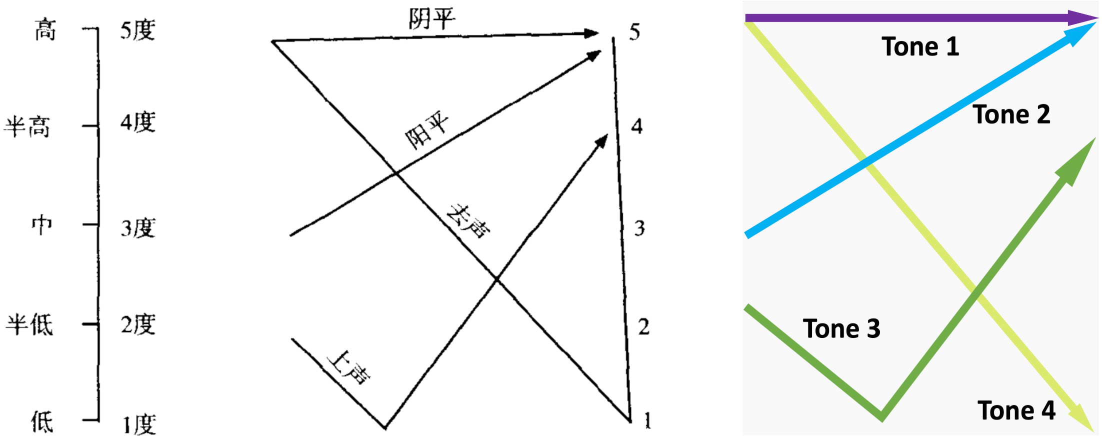


Figure 2.2 Schematic contours of the four lexical tones in Standard Mandarin (right). The figure on the left was from the handbook Xian Dai Han Yu ("Modern Chinese", Wang et al., 2006).

2.1.3.2 Mandarin dialect tone systems

All Mandarin dialects are tonal; each morphemic syllable carries a specific tone category. Most Mandarin dialects have a four-tone system as Standard Mandarin and a few have three or five tones (Cheng, 1991; Ho, 2003). Dialect dictionaries have been using Chao tone numerals to schematically illustrate pitch contours of the lexical tones in Mandarin dialects. Table 2.5 lists out the recorded tone systems of the six Mandarin dialects under investigation from two canonical references: *Xiandai Hanyu Fangyan Dacidian* (2002) ("*The Modern Dictionary of Chinese Dialects"* as Ref. 1) and *Hanyu Fangyin Zihui* (1989) ("*The Sounds of Chinese Dialects"* as Ref. 2). Beijing, Chengdu, Jinan, Wuhan and Xi'an Mandarin each have four lexical tone categories; Taiyuan has a three-tone system with Tone 1 and Tone 2 merged. The Standard Mandarin tone system was added in Table 2.5 for reference.

Table 2.5 Tone systems of six Mandarin dialects in Chao tone notation according to the two references published in 2002 (Ref. 1) and 1989 (Ref. 2). Bolded numbers correspond to a discrepancy between the two sources.

	sources	Tone 1	Tone 2	Tone 3	Tone 4
Standard Mandarin	Table 2.4	55	35	214	51
Beijing	Ref. 1	55	35	214	51
	Ref. 2	55	35	214	51
Chengdu	Ref. 1	**55**	**21**	53	**213**
	Ref. 2	**44**	**31**	53	**13**
Jinan	Ref. 1	213	42	55	21
	Ref. 2	213	42	55	21
Taiyuan	Ref. 1	11		53	45
	Ref. 2	11		53	45
Wuhan	Ref. 1	55	213	42	35
	Ref. 2	55	213	42	35
Xi'an	Ref. 1	21	24	53	**44**
	Ref. 2	21	24	53	**55**

The Chao-tone-based descriptions were initially compiled based on the impressionistic transcriptions by trained linguists at the time and have been the important reference for studies on the tone systems of Mandarin dialects. In the past two decades, more researchers have started collecting speech data from present-day speech communities and applying sophisticated acoustic-phonetic analyses to validate the previous descriptions. Empirical research has also steered from concentrating on Standard Mandarin or Beijing Mandarin to investigating dialectal variation of the sound systems on both segmental and suprasegmental levels. For tone systems specifically, the tone inventories of several Mandarin dialects have been thoroughly analysed, while many of others remain under-investigated. The following part reviews the acoustic-phonetic studies on the tone systems of a few Mandarin dialects relevant to the current chapter.

Chao tone descriptions of the Beijing tone categories are the same as Standard Mandarin. Specifically, Shi and Wang (2006) measured stepwise F0 values for each tone category in Beijing Mandarin using recordings from fifty-two native speakers. The F0 values were converted to t-values within the 0–5 range. Figure 2.3 demonstrates the measured pitch contours for Beijing Mandarin. Beijing Tone 1 was phonetically realised between [44] and [55] and showed a gradual decline in F0 towards the end of the monosyllables. For Beijing Tone 2, the average start point was 2.4; the lowest point was around 2.1 at the 30% of the duration; the highest pitch point reached 4.02. Shi and Wang (2006) specified possible descriptions of Tone 2— [435], [335] and [224] and considered them as legit variations for the [35] description. Tone 3 in Beijing Mandarin occupied the lower range of the pitch values: the tone started from 2.19 and reached the lowest level around 40% into the vowel before rising to the highest point at 2.51. Tone 3 can be transcribed as [313], [212], [314] or even [214] considering the difference between the start and the end points. Beijing Tone 4 followed a downward trend and was described as [51] or [52].

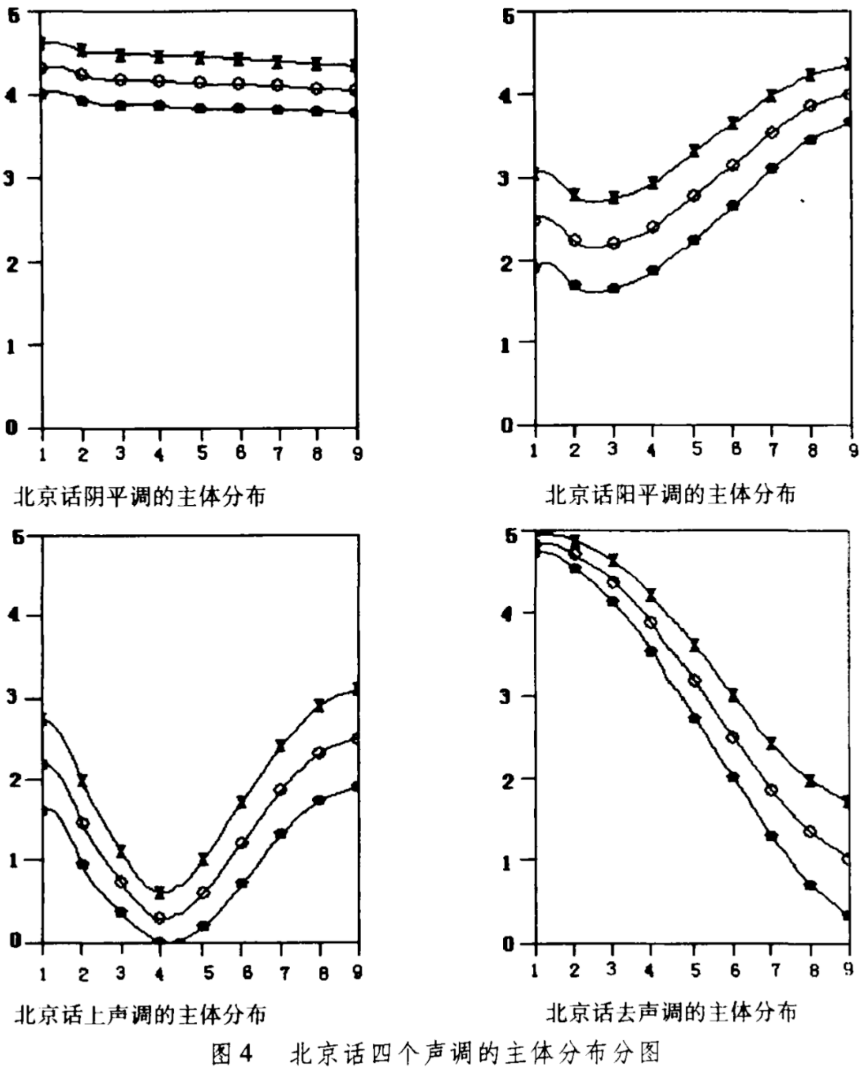
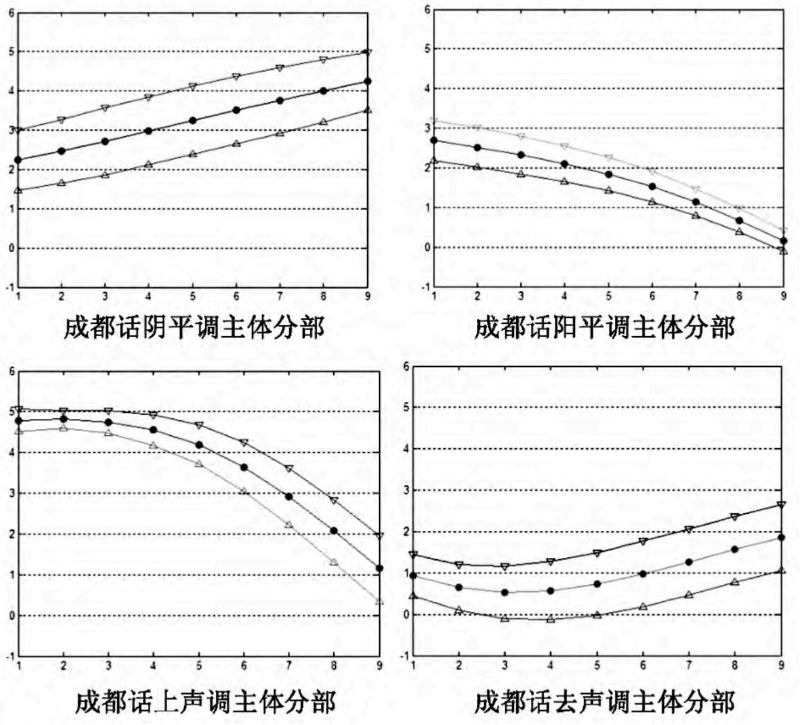


Figure 2.3 The measured tone contours of Beijing Mandarin by Shi & Wang (2006). The original figure was Fig. 4 in the paper. Each tone was represented by three lines: the middle stood for the mean values; the upper and the lower lines represented one standard deviation from the mean.

He (2015) adopted the same approach used in Shi and Wang's (2006) study to examine the phonetic realisation of four lexical tones in Chengdu Mandarin. The recordings from thirty-eight native speakers were used for analysis. Figure 2.4 presents the pitch trajectories for each tone category. The Chengdu tone system was described as: Tone 1 [45], Tone 2 [21], Tone 3 [42], and Tone 4 [213]. He's measurement of Tone 1 as a rising tone was different from the previous record, where Tone 1 was considered a high level tone (Table 2.5). Chengdu Tone 2 and Tone 4 primarily differed in the onset pitch level: the mean start point was 2.68 for Tone 2 and 4.67 for Tone 3. Moreover, Chengdu Tone 3 maintained the onset pitch till the 30% of the duration and even rosed up a bit around the 20% of the vowel production. Chengdu Tone 4 showed a dipping pattern which was transcribed as [212] with the alternatives being [223], [112] and more likely [213].



成都话阴平调主体分部

成都话阳平调主体分部

成都话上声调主体分部

成都话去声调主体分部

图3　成都话四个声调的主体分布分图

Figure 2.4　　The measured tone contours of Chengdu Mandarin by He (2015). The original figure was Fig. 3 in the paper. Each tone was represented by three lines: the middle stood for the mean values; the upper and the lower lines represented one standard deviation from the mean.

Research on the tonal inventories of Mandarin dialects was done in comparison with Standard Mandarin. For instance, Liu, Chen and Schiller (2020) empirically tested the mappings between the lexical tones in Xi’an Mandarin and Standard Mandarin, and found that the tones with the same contour type between Xi’an and Standard Mandarin were to a large extent perceived as the same. Thirty bidialectal speakers of Standard Mandarin and Xi’an Mandarin contributed to the speech data. Figure 2.5 shows the Xi’an tone contours based on the acoustic-phonetic measurements (Liu et al., 2020). The Xi’an Mandarin tone system was transcribed as: Tone 1 [21] or [31], Tone 2 [24], Tone 3 [52], [53] or [42], and Tone 4 [44], [55] or [45]. Additionally, Wu et al.’s (2016) study compared the tone systems of Jinan Mandarin and Standard Mandarin with a focus on the systematic relation between Mandarin dialects (see more in the following section).

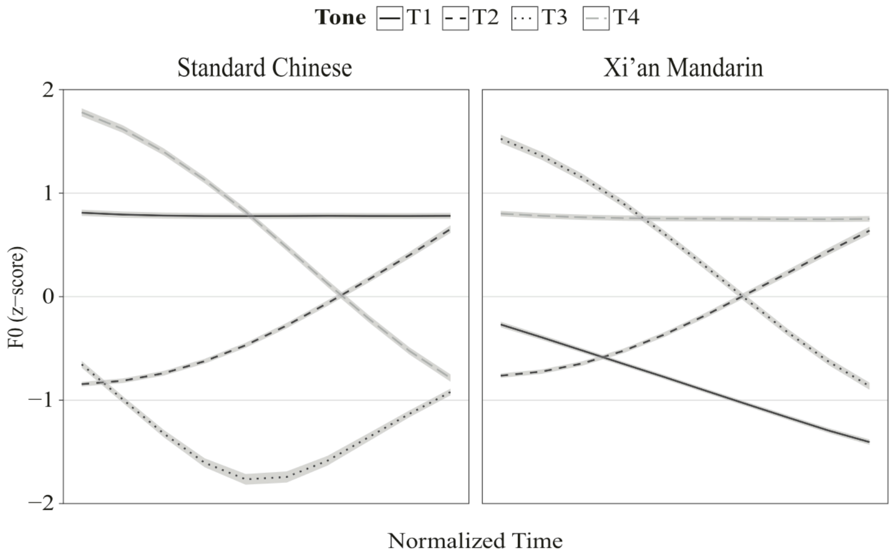


FIG. 1. Mean F0 (*Z*-score) contours of the four tones in SC and XM. The F0 values of each tone were averaged over 30 speakers and 30 monosyllabic items with the tone of each item represented by 10 equally distanced F0 values taken from the rhyme part of the time-normalized item. The grey areas indicate the 95% confidence interval of the corresponding mean.

Figure 2.5 The measured tone contours of Standard Mandarin and Xi’an Mandarin in Liu et al. (2020). The original figure was FIG.1 in the paper.

2.1.3.3 Systematic word–tone mappings across Mandarin dialects

For native speakers of Standard Mandarin, Tone 1, Tone 2, Tone 3 and Tone 4 are used as the arbitrary labels for the four phonological categories, whereas for speakers of other Mandarin dialects, the dialectal tone categories might not perfectly align with those of Standard Mandarin, especially for dialects that do not have exactly four categories such as Taiyuan Mandarin which has a three-tone system. However, previous research has suggested that, at least for the dialects with the same number of tone categories, a systematic word–tone relationship exists between Mandarin dialects and Standard Mandarin, which means that words realised in one tone category in one dialect are most likely still realised as one category in another dialect and these words mostly likely belong to the same phonological tone category. To test whether there is a systematic relation of tone categories between Standard Mandarin and a Mandarin dialect, Wu et al.'s (2016) experimented on Jinan Mandarin and successfully predicted tone realisations of Jinan from the corresponding tone categories of Standard Mandarin. This relatively consistent word–tone relationship between Mandarin dialects makes it feasible to comparatively analyse the dialectal tone systems assuming one set of phonological tone categories. Nevertheless, a sanity check would be needed to ensure identification of a possible tone merger or splitting, so that the analysis of the tone categories could be modified accordingly.

Though the current study uses a phonological categorisation of the tone system, an alternative analysis of the tone acoustics is to use computational clustering directly on the acoustic measurements without reference to the phonology. For example, Wu, Chen, van Heuven and Schiller (2018) examined the phonetic tones of Jinan Mandarin, using functional partitioning to automatically compute clusters of tone realisations, and eventually generated eleven different pitch contours. This method seems capable of stratifying phonetic tone categories as accurately as possible; however, without presumptions of the phonological categories, clustering results might indicate an over-complicated tone system with an unrealistic number of tone categories. Given the typological studies on tone systems of Mandarin dialects, it is highly improbable that the number of tone categories of Mandarin dialects exceeds the number of six (Szeto et al., 2018). Moreover, computationally clustered tone categories might not be the same thing as the perceived tone categories. Therefore, this study took the phonological category approach and used the four tones of Standard Mandarin as the pre-assigned categories, to which the other Mandarin dialects were then compared.

2.1.4 Acoustic parameters of lexical tones

The production study in this chapter aims to conduct an acoustic-phonetic analysis of the lexical tone systems of Standard Mandarin and the six Mandarin dialects. F0 height and contour have been considered the primary acoustic cues of Mandarin lexical tones (Ho, 1976; Duanmu, 1999; Wang, Jongman, & Sereno, 2003; Jongman, Wang, Moore, & Sereno, 2006; Wang, 2013; Zhang, 2018). Previous studies also reported on the relative contribution of F0 height and F0 contour in tone identification tasks (Gandour, 1984; Zhang & Kirby, 2020), but the results are contingent on the specific dialect and tone categories being tested in the experiment. Gandour (1984) reported greater importance of F0 height in tone identification for Cantonese speakers than Mandarin and Taiwanese speakers. Zhang and Kirby (2020) found that between Cantonese Tone 4 (low-falling tone) and Tone 6 (low-level tone), F0 height was more critical than F0 contour in Cantonese low tone perception.

Two other F0-related measures are also useful to capture finer F0 characteristics—the turning point and ΔF0 (Moore & Jongman, 1997; Wang et al., 2003). The turning point refers to the time at which the contour alters its trend from falling to rising, or from rising to falling; ΔF0 is the difference in F0 from the onset to the turning point. In Standard Mandarin, for example, both Tone 2 and Tone 3 show a dipping pattern phonetically, but Tone 3 typically has a later turning point, about 41% of the average duration and a much larger ΔF0 than Tone 2; Tone 2 is most likely perceived as rising, instead of dipping despite a dip at 15% of the mean syllable duration (Ho, 1976). Perception studies of Tone 2 and Tone 3 also show that the turning point as a temporal aspect could function as a primary cue in Tone 2 vs Tone 3 distinction (Moore & Jongman, 1997; Wang et al., 2003). For the comparison between Tone 3 and Tone 4, Tone 4 most often does not show a clear turning point in F0 values as it starts at a high value and falls sharply to a low value.

Apart from F0 measures, Mandarin tones may also differ in their overall duration (Howie, 1970, 1976; Chuang & Hiki, 1972; Ho, 1976). Standard Mandarin Tone 2 and Tone 3 generally have longer durations than Tone 1 and Tone 4. In addition, intensity differences of lexical tones seemed randomly distributed and varied depending on the position in a phrase or a sentence; they also have minimal impact on tone perception (Jongman et al., 2006).

2.1.5 Quantification of tone system similarity

A primary question in investigating the vowel spaces and lexical tones of Mandarin dialects is to understand to what extent they differ from each other along the acoustic-phonetic dimensions. Based on previous descriptions, we expect comparable segmental systems and distinct phonetic tone inventories of the six Mandarin dialects. Few empirical studies have directly used acoustic data to investigate the similarity or dissimilarity of the Mandarin dialect tone systems. Instead, dialect similarity was primarily assessed by intelligibility tests and edit distance based on phonological transcription of the sound system in the previous studies. Specifically, Tang & van Heuven (2007, 2008, 2009, 2011) conducted a series of studies on mutual intelligibility of fifteen Chinese dialects in the semantic identification tasks, where participants were instructed to identify the semantic category of the heard target word in the sentence. Table 2.6 summarises the intelligibility scores for Beijing listeners when presented with the six Mandarin dialects (Tang & van Heuven, 2009); higher scores indicated higher intelligibility level. According to Tang's results, native speakers of Beijing Mandarin understood Jinan and Taiyuan Mandarin (77) better than Hankou (67) and Chengdu Mandarin (62); Xi'an Mandarin (58) was least intelligible within the comparisons. Hankou Mandarin is a major local variety of Wuhan Mandarin.

Table 2.6 Intelligibility percentage scores for Beijing listeners presented with the six dialects.

listener	*speaker*					
	Beijing	Chengdu	Jinan	Taiyuan	Hankou (Wuhan)	Xi'an
Beijing	98	62	77	77	67	58

Apart from intelligibility tests, edit distance, i.e. *Levenshtein distance*, was used for quantifying between-category similarity based on the phonological transcriptions. *Levenshtein distance* counts the minimum number of edits to transform one string in a variety to its corresponding string in another variety. There are three basic edit types—deletion, insertion and substitution. By default, deletion and insertion count as 0.5 edit point and substitution counts as 1 point; however, different point-assigning schemes can be employed given different weighting operations. *Levenshtein distance* was extended to measure lexical tone distance as well. Chao tone numerals of each tone category are treated as discrete digits and each digit counts as a single string aligning from the left (Tang & Heuven, 2011). For example, between the tone [51] and tone [35], there are two comparable strings—5 vs. 3 and 1 vs. 5; to transfer from [51] to [35] it takes two edits, both as substitution.

To measure edit distance, text/digit length needs to be normalised through dividing the counted distance by the number of alignment slots (Gooskens, Heeringa & Beijering, 2008). The number of alignment slots equals the number of the longest slots of the pair; the normalised edit distance is a score ranging from zero to one. Table 2.7 (a) is an example of calculating edit distance between the phonetic tone [55] and [241] of the same phonological category in two dialects. A major drawback of this approach is that it does not calculate the phonetic variation along each slot; for lexical tones, changes in the relative pitch level are unitarily considered as substitutions regardless of the actual pitch difference. A possible revision for calculating tone distance is to count the absolute difference between each paired Chao tone numerals; empty slots are counted as 0 (Figure 2.7 b). The revised form might better inform between-dialect tonal difference than the previous method.

Table 2.7 Example of non-normalised and normalised edit distance.

(a) Non-normalised edit distance

Chao tone transcription	Dialect A	5	5	
	Dialect B	2	4	1
String operation		Sub.	Sub.	Ins.
Points/costs		1	1	0.5
The number of longest alignment slots		3		
Non-normalized edit distance		2.5		
Normalized edit distance		0.833 (2.5/3)		

(b) Normalised edit distance

Chao tone transcription	Dialect A	5	5	(0)
	Dialect B	2	4	1
Absolute point distance		|3|	|1|	|-1|
The number of longest alignment slots		3		
Non-normalized edit distance		5		
Normalized edit distance		1.667 (5/3)		

2.1.6 Speech data collection

Speech science has traditionally relied on high quality speech recordings, collected in person with a standardised recording setup in laboratories/studios or in field settings. The strictly controlled environment for speech recording allows the acoustic-phonetic analysis to be focused on differences in the speech alone. However, recent global developments ranging from COVID-19 to climate change have triggered a comprehensive re-evaluation of our approach to research and travel, while researchers have drastically rethought data collection approaches and turned to varieties of remote speech data collection. Remote audio collection hereby refers to the scenario where the researcher delivers an experiment to a participant virtually and remotely, who in turn records his/her speech using a personally available recording device, such as a smartphone or computer, and sends the recording back to the experimenter.

Early attempts at remote audio collection were inevitably confined to the state of technological development at the time. Previous methods in the 1990s exploited access to landline telephones to collect speech corpora (e.g. CALLHOME; Canavan and Zipperlen, 1996). This method increased the range of participant enrolment, but the recordings were limited by the telephone bandwidth, and were thus of low quality for phonetic research. Thanks to current smartphone technology, high quality recording baselines and its widespread accessibility can advance remote audio collection for research in speech sciences and engineering. As approximately 50% of the world's population is estimated to own a smartphone and this figure is likely to rise, remote audio collection with smartphone devices can increase both the amount and range of speech data. With a robust method, we can simultaneously reach a vast number of diverse communities around the world: anyone with access to a smartphone could contribute. Speech technologies such as speech-to-text systems and text-to-speech synthesis might also benefit from large quantities of data which allows for more precise estimates of true population parameters.

Moving forward with any type of remote data collection requires reasonable control of the potential variability introduced by the recording environment, recording devices and uncertainty in implementation (Leemann et al., 2020). Previous work (De Decker & Nycz, 2011) tested the consistency of recording devices within a laboratory but on older technology (e.g., 2010-era mobile phones and laptops). Bird et al. (2014) investigated the use of smartphones for recording in isolated indigenous communities, but focused on the feasibility of general speech collection as opposed to between-device or environmental effects on acoustic-phonetic measures. More recently, Grillo et al. (2016) identified significant variation between smartphone devices in measurements of voice quality in a laboratory setting. Work from Leemann et al. (2020) has also investigated remote speech collection via a smartphone application, but in the same country with simultaneous videoconferencing. In contrast, the audio recordings collected in this study were from a culturally and geographically remote region (UK to China), without videoconferencing. More specifically, Zhang et al. (2021) tested and compared the reliability of the acoustic data collected using Zoom application and smartphone recording applications, including Awesome Voice Recorder (AVR, Newkline, 2020) and Recorder (DawnDIY, 2016). Their results suggested equally accurate measurement of F0 for all the tested devices and better formant measures using phone applications than Zoom.

Speech resources for Mandarin Chinese have been substantially expanded over the past few decades through telephone conversation collection (e.g. CALLFRIEND, Canavan and Zipperlen, 1996) and lab recording (e.g. ALLSSTAR, Bradlow, n.d.), but most related corpora have primarily collected speech data in Standard Mandarin and only a few contained regional varieties as part of the research (e.g. NCCU, Chui and Lai, 2008). Although regional dialects of the Mandarin branch (in contrast with the non-Mandarin branch, such as Wu or Cantonese) are spoken by a large proportion of the population in mainland China, recordings of these dialects that are available for speech science research remain scarce.

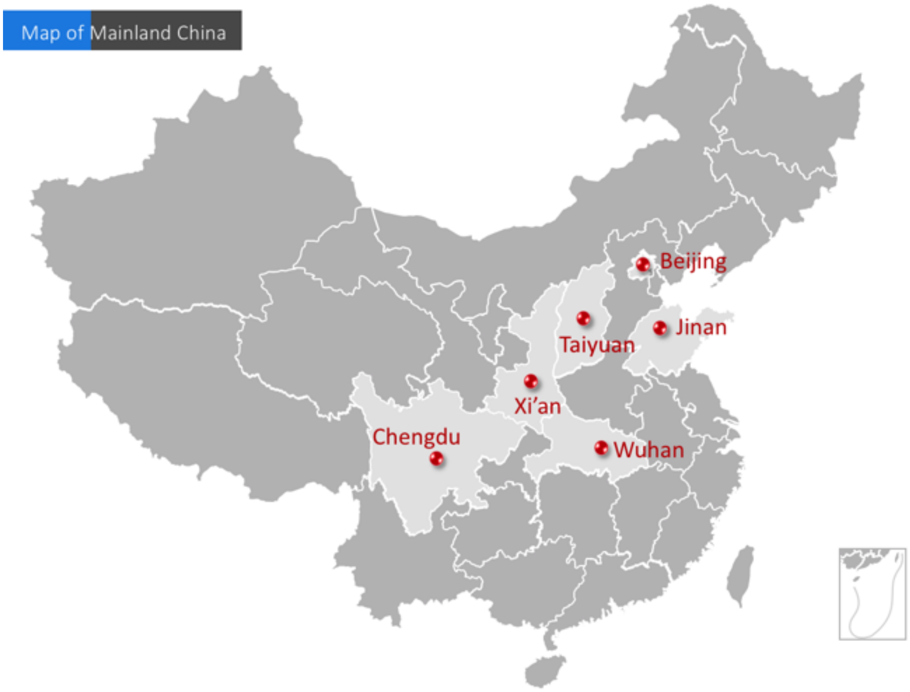


Figure 2.6 Locations of cities where the six Mandarin dialects are spoken (Zhao & Chodroff, 2022).

2.1.7 The present chapter

With remotely collected speech data, this chapter aims to 1) build a multi-speaker spoken corpus of six Mandarin dialects—Beijing (BEI), Chengdu (CHD), Jinan (JNN), Taiyuan (TYN), Wuhan (WHN) and Xi'an (XIA) (Figure 2.6 demonstrates the approximate locations of cities where the six dialects are spoken); 2) conduct an acoustic-phonetic analysis on vowel spaces and tone inventories of the six dialects; comparable vowel spaces and distinct tone systems are expected across dialects; 3) collect speech data as the stimuli for the perception experiments in Chapters 3 and 4. The speech data included the recordings of monosyllabic words, disyllabic words, short sentences, the North Wind and Sun passage, and the Wo Chun poem. All collected speech data have been made available on OSF at https://osf.io/fgv4w/. The present analysis then makes use of the monosyllabic words, as production of lexical tones on monosyllabic words avoids tonal variation due to contextual influence, which is particularly useful for mapping out the tone system of a given dialect (Xu, 1997).

2.2 Methods

A speech production experiment was created and hosted online using the Gorilla Experiment Builder (Anwyl-Irvine, Massonié, Flitton, Kirkham & Evershed, 2020). Speech data were collected between August and October 2020. The participants were asked to read the given material presented through the Gorilla web-based project in both Standard Mandarin and their regional dialect that was one of the six dialects—Beijing, Chengdu, Jinan, Taiyuan, Wuhan, and Xi'an dialect. They simultaneously recorded themselves using the designated smartphone recording application and then uploaded recordings to a cloud drive folder upon finishing all the tasks.

2.2.1 Subjects

Thirty-six speakers have thus far been recruited for the corpus. The participants were native speakers of one of the six dialects (Table 2.8), aged between 18 and 45, proficient in Standard Mandarin and literate in the Standard Simplified Chinese writing system with no auditory impairment or reading difficulty. The participants were expected to be digitally literate to complete the online experiment.

Table 2.8 Participants in the speech production experiment (F for female, M for male).

Beijing Mandarin	9	(F: 6, M: 3)
Chengdu Mandarin	5	(F: 4, M: 1)
Jinan Mandarin	5	(F: 3, M: 2)
Taiyuan Mandarin	7	(F: 4, M: 3)
Wuhan Mandarin	6	(F: 4, M: 2)
Xi'an Mandarin	4	(F: 2, M: 2)
Total	36	

2.2.2 Materials

Five sets of reading materials were constructed for the study: monosyllabic words (Word List 1), disyllabic words (Word List 2), short sentences, the North Wind and the Sun passage, and a modern Chinese poem, Wo Chun ("卧春"). All the reading materials were attached in the Appendix. Each type of material was read first in Standard Mandarin and then in the participant's dialect. Standard Mandarin speech was recorded first to familiarise participants with the content. There were altogether ten tasks for the five sets of reading material per participant.

The list of monosyllabic words consisted of 40 unique characters that were constructed by pairing ten unique monosyllables each with one of the four lexical tone categories in Standard Mandarin (10 syllables × 4 tone categories). The chosen monosyllables were the ones that allow realisations of all four tone categories. For example, the syllable "ma" (in Pinyin) could be realized as "ma1", "ma2", "ma3", "ma4" (numbers for the tone categories). The neutral tone was excluded in this task as it does not typically occur in monosyllabic words in isolation. Four of the syllables had a CVC template, four with a CV template and two with a V template.

Twenty disyllabic words were included in the disyllabic word list. Each word represented a unique combination of lexical tones over two syllables. For disyllabic words, the second syllable can optionally be assigned a neutral tone, so there were altogether 20 tone combinations (4 tone categories for the first syllable × 5 tone categories for the second syllable). Again, all the syllables in the disyllabic words allowed a full set of tone realisations in Standard Mandarin.

The short sentences were originally designed as the stimuli for the perception experiments on the effect of semantic plausibility on tone perception (see more in Chapters 3 and 4). There were twenty-four pairs of sentences each containing a target word at either sentence-final (12 pairs of sentences) or sentence-medial position (12 pairs of sentences). Each pair of sentences shared the same sentence structure with the only difference in the tone realisation of one target word. Specifically, one sentence from the pair was grammatically correct and semantically plausible, while the other sentence altered the tone category of the target word (sentence-final or sentence-medial), thus making the sentence semantically implausible. Overall, half of the sentences were semantically plausible, and half were semantically implausible.

The passage material in the experiment was *the North Wind and the Sun* story. This was to provide recordings of the six dialects in connected speech. The Standard Mandarin translation of the story was used for recording speech of both Standard Mandarin and regional dialects as the words in Standard Mandarin are to a large extent shared in the lexicon of those dialects. One native speaker of each dialect helped to check the materials before the experiment.

The poem was adapted from a modern poem by Chinese novelist, Han Han (2000). All the characters used in the poem were homophonic in Standard Mandarin: by the written form, the poem depicted a tranquil scene of early spring, but if read aloud with the same segmental sequence and a different tonal sequence, the poem could be perceived as a monologue of a man mocking himself for being silly. The original spring scene version of the poem was recorded in both Standard Mandarin and the regional dialects as potential stimuli for future studies on speech perception. The recordings of the poem were collected for future study, hoping to follow up on Zhao's degree project on the activation of phonological information during silent reading (Zhao, 2017).

2.2.3 Procedure

Once recruited, the participants received a web link to the experiment. After a brief introduction to the study and presentation of the consent form, participants were given a detailed set of instructions for completing the experiment and at-home recording using smartphone devices. It was recommended that they use an extra device, such as a desktop, laptop, or tablet to display the Gorilla page with the reading materials and meanwhile use a smartphone to make the recordings.

For preparation, the participants attended to specific requirements of the recording environment and operational instructions. If possible, the recording was to be made in a quiet room with plenty of soft furnishing to reduce reverberation, such as a bedroom, and with the doors and windows closed. The recording device, the participant's own smartphone, was to be placed on a hard surface, such as a table or a desk for better recording quality. It was also suggested that participants maintain a distance of approximately 20 to 30 cm from the microphone and to keep this distance throughout the experiment.

To prepare the recording device, participants installed a freely available sound recording app on their smartphones. iPhone users installed the app called Awesome Voice Recorder by Newkline Co., Ltd. with version 8.0.4 or later. Users of Android devices installed the app called ASR recorder by NLL APPS. Participants were then instructed to open the "settings" menu from AVR on iPhone or ASR on Android device, and set the following entries as specified: file format as WAV files, encoding quality at medium, sample rate at 44100 Hz, bit rate at 128kbps, and channel as stereo. (These were later converted to mono recordings.) All other settings were left in their default state. These two apps were chosen because of their similar setting options so that the settings for both apps were kept identical.

In addition, a unique and anonymous experiment ID was generated for each participant to link the recording to the transcript from Gorilla. This was done to minimize the use of any identifiable information in the file naming convention.

In the practice trials, the participants were asked to read aloud an additional list of 10 disyllabic words presented in characters in Standard Mandarin, record themselves, check the recording quality and contact the experimenter if needed. The practice trials were not included in the final corpus. After practice, the participants started recording in sequence—Word List 1, Word List 2, sentences, the passage, and the poem. For each type of material, they were instructed to read the given material first in Standard Mandarin as one task and then in their regional dialect as another task for a total of ten tasks.

All items were presented in characters at the centre of the webpage. For the word lists and sentences, participants proceeded through each item by clicking the "next" button on the screen. These items were fully randomized. For the passage and poem, the text was presented on one screen. The participants were instructed to make an individual recording for each task so that the size of each file was easy to upload and send. At the start of each task, participants began a new recording, and then read aloud each item. By the end of the task, the participants saved the recording and named the file as "experiment-ID_task-number" (e.g. lz0916xlh_01). The task number was provided at the end of each task.

After completing all recording tasks, the participant uploaded the ten WAV files by scanning a QR code that directed to a Tencent (Weiyun) cloud folder. Files could then be uploaded anonymously and were accessible only to the experimenter. In addition to the recording files, the participants also uploaded a special QR payment code generated from WeChat Pay—a secure payment app widely used in China and similar in function to PayPal. We used this QR payment code to make a payment to the participants. For a full recording (completion of all tasks and expected quality of the recording), the participant was compensated with ¥45 CNY (approximately £5). For receipt of any partial recording (partial completion of the task or poor recording quality due to technical issues), the participant received ¥30 CNY (approx. £2.70). Thirty-five participants thus far have provided full recordings; one participant provided nine recordings out of ten. The uploaded recordings were then downloaded, immediately de-identified, and transferred to a secure folder on a password-protected computer owned by the university.

A post-experiment survey was then conducted to record the participants' dialectal background, demographic information, devices used for recording and general feedback on the experiment. For dialectal background, we recorded the participants' regional dialect to include both city- and county-level information; the participants also rated their fluency and proficiency in speaking Standard Mandarin and the regional dialect, and how frequently the regional dialect was used. For demographic information, the participants reported gender, age, and educational level. Information on the recording device included device type (mostly

Huawei and iPhone) and how long the device has been in usage. Feedback questions were presented as rating scales for the participant's opinion on the experimental design, plus one optional text box for suggestions. In addition, Gorilla automatically collected information on local timestamps and the devices that were used to open the web link, device information including device type, operating system and browser.

2.2.4 Corpus Annotation

2.2.4.1 File Naming

Before annotation, the collected recordings were renamed with the following format: native dialect, speaker sequential number, gender, recorded dialect, and task, each separated by an underscore. The native dialect codes were BEI for Beijing dialect, CHD for Chengdu, JNN for Jinan, TYN for Taiyuan, WHN for Wuhan, and XIA for Xi'an. The speaker sequential number was a three-digit code automatically generated by Gorilla in the temporal order of participation in the experiment, e.g. the first participant was 001. Gender was coded as F for female speakers and M for male speakers. The recorded dialect referred to whether the material was recorded in Standard Mandarin, coded as CMN, or the participant's own dialect, coded the same way as for the native dialect. Tasks were coded as WL1 for the monosyllabic word list, WL2 for the disyllabic word list, SST for short sentences, NWS for the *North Wind and the Sun* story, and WCH for the *Wo Chun* poem.

An example file name with the .wav extension is as follows: CHD_012_F_CMN_WL2.wav, meaning this is a recording of Word List 2 (disyllabic words) spoken in the Standard Mandarin dialect by a female native Chengdu dialect speaker who was the 12th participant in the experiment.

2.2.4.2 Transcripts

Initial transcripts per speaker per task were accessed directly from *Gorilla* and then processed with RStudio to generate text-format transcripts that matched the content of each recording. A total of 317 transcripts were obtained from Gorilla. A few transcripts were missing or did not match the content of the recording. This data loss arose from several noticeable issues. If a participant restarted the experiment at any point, then Gorilla regenerated transcripts for all tasks, meaning any previously completed task was overwritten. We also speculated that some data loss was due to internet instability during the experiment or other potential technical errors from Gorilla.

2.2.4.3 Forced Alignment

All recordings were force aligned at the utterance-, word- and phone-levels using a combination of Praat (Boersma and Weenink, 2020) and the Montreal Forced Aligner (MFA) (McAuliffe, et al., 2017).

The recording-level transcripts were first aligned to the audio at the utterance level using a custom-made Praat script and then manually checked. All recordings were aligned using the pretrained Mandarin acoustic model released with the MFA. This model was trained on the Mandarin subset of the GlobalPhone corpus. Mandarin pronunciation dictionaries were created using the corresponding pretrained Mandarin G2P model, also released with the MFA. These were then manually corrected by a native speaker. The dialect recordings were aligned using MFA v1.0 and the Standard Mandarin recordings using MFA v2.0.0b9. Figure 2.7 shows an example output TextGrid with word- and phone-level annotations from the MFA along with its corresponding WAV file.

Misaligned sonorant boundaries were manually checked and corrected for the recordings of monosyllabic words (Word List 1). Additional manual corrections of the TextGrid alignments will be uploaded and documented on the OSF website.

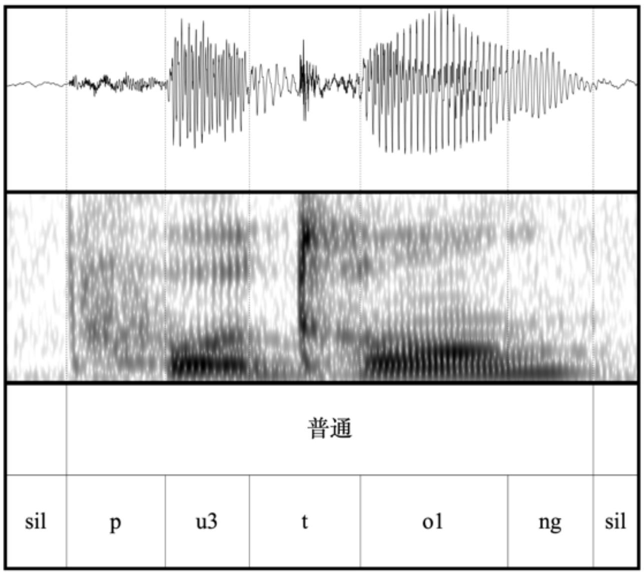


Figure 2.7 Part of the WAV file for the disyllabic word “普通” <pu3 tong1> and its corresponding TextGrid in Praat.

2.2.5 Acoustic measurement

2.2.5.1 Formant measurement

For vowel spaces, mid-point F1 and F2 and duration of the vocalic part were measured on the four monophthongs /i, a, u, ə/ (“i, a, u, e” in Pinyin) automatically using a Praat script. Mean F1 and F2 values were first calculated by speaker and vowel. Outliers that fell out of two standard deviations to the speaker-vowel-specific mean formant were excluded. The speaker-intrinsic variation was normalised using Lobanov’s method (Lobanov, 1971; Nearey, 1977;

Adank, Smits & van Hout, 2004; Flynn, 2011; Wissing & Pienaar, 2014). F_n is the individual formant measurement to be normalised; $mean\_n$ is the mean value of the formant n for the speaker; SD_n is the standard deviation of the formant n for that speaker. The formula used is: (n = 1 for F1, n = 2 for F2)

$$F_{n\_normalised} = \frac{(F_n - mean_n)}{SD_n}$$

2.2.5.2 F0 measurement

As phonetic tone realisations are largely affected by the adjacent tonal contexts in terms of anticipatory effects and carry-over effects (Xu, 1994, 1997), recordings of the monosyllabic word list were used for the tonal analysis. F0 contours were used to represent the phonetic realisation of the lexical tone (Jongman et al., 2006; Tupper et al., 2020). Ten equally spaced F0 values in hertz were automatically extracted over the sonorant portion of the word; F0 at the onset of the sonorant part was also extracted for reference. The extraction range was 75 – 500 Hz for female speakers and 75 – 250 Hz for male speakers in Praat. F0 values were converted to semitones in the first place to normalise between-speaker variability for each dialect. Chao tone conversion was further applied to normalise both speaker and dialect difference within a comparable range.

To be specific, semitones were calculated with the following formula (Yuan & Liberman, 2014), where F0_base refers to the speaker-specific F0 value in the 5th percentile. For each of the extraction points, the by-speaker mean F0 was calculated; the grand mean and standard error were then derived over speakers in each dialect group.

$$\text{Semitone} = 12 \times log_2(\frac{F0}{F0\_base})$$

In addition to the semitone conversion, the measured F0 values were fitted in a Chao tone conversion using the t-value formula proposed by Shi (2008), also used by Jóźwik & Shi (2018):

$$Chao\ tone = 5 \times \frac{log_{10}\left(F0_{n\_mean}\right) - log_{10}(F0_{speaker\_min})}{log_{10}\left(F0_{speaker\_max}\right) - log_{10}(F0_{speaker\_min})}$$

$F0_{n\_mean}$ is the speaker's mean F0 value for a given extraction point (n = 0~10 in this study) for a tone category. $F0_{speaker\_min}$ is the speaker's minimal average pitch and $F0_{speaker\_max}$ is the speaker's maximal average F0. This formula was used to normalise between-speaker and between-dialect variability within the same 0–5 scale so that comparisons can be made across dialects concerning the relative pitch height and contour, regardless of the absolute pitch values, age, gender among other factors.

2.3 Results and discussion

The results comprise two main sections: Section 2.3.1 for dialectal vowel space analysis and Section 2.3.2 for the acoustic-phonetic analysis on the tone inventories. The tonal analysis (Section 2.3.2) was conducted in four parts: the reliability of the measured tone acoustics was first validated using the production data of Standard Mandarin (Section 2.3.2.1), which matched existing documentation of the pitch contours; dialect-specific tone inventories were illustrated and discussed in Section 2.3.2.2; individual tones of each contour pattern were further compared and analysed across dialects (Section 2.3.2.3); as an attempt to quantify dialectal similarity of the tone system, edit-distance analysis was applied using the updated Chao tone numerals based on the measured acoustics (Section 2.3.2.4).

2.3.1 Vowel spaces

Normalised mean F1 and F2 values by vowel are plotted in Figure 2.8. Figure 2.9 presents the scatterplots of normalised F1 and F2 values by dialect and vowel. Variation in normalised F1 and F2 values was analysed respectively using linear mixed-effects models with fixed effects of vowel, dialect, gender and the interaction between vowel and dialect, and a random intercept for speaker. Sum coding was used for vowel and dialect to compare each level to the overall mean of the dependent variable; gender factor was sum-coded (gender: female = 1, male = −1).

For F1 values, significant main effects emerged for vowel conditions ([i]: β = −0.84, t = −27.13; [u]: β = −0.50, t = 12.74; [a]: β = 1.4, t = 45.27), but not for dialect conditions (*Chengdu*: β = 0.0030, t = 0.0072; Jinan: β = 0.0273, t = 0.648; Taiyuan: β = 0.0053, t = 0.103; Xi’an: β = −0.0173, t = −0.401; Wuhan: β = 0.0236, t = 0.580), which indicated that [i, u, a] occupied disparate positions in terms of tongue height compared to the average, but no significant difference was found across dialects for F1 values. The interactions between vowel and dialects did not reach significance except for [u] (β = 0.2137, t= 2.080) and [a] (β = −0.2175, t = −2.658) in Taiyuan, and [i] (β = 0.1333, t = 2.028) in Wuhan. Specifically, [u] in Taiyuan was produced with significantly higher F1, while [a] was realised with significantly lower F1 than the average. For F2 values, the model revealed a significant main effect of vowel ([i]: β = 1.0938, t = 23.266; [u]: β = −0.8480, t = −14.266; [a]: β = −0.1811, t = −3.830) in the expected directions, but no significant effect of dialect (Chengdu: β = −0.0424, t = −0.688; Jinan: β = −0.0493, t = −0.766; Taiyuan: β = 0.0473, t = 0.610; Xi' an: β = −0.0117, t = −0.176; Wuhan: β = 0.0133, t = 0.214) or the interactions between dialect and vowel. These findings indicate significant overall differences between [i, u, a] in F2 values, but no significant differences across dialects.

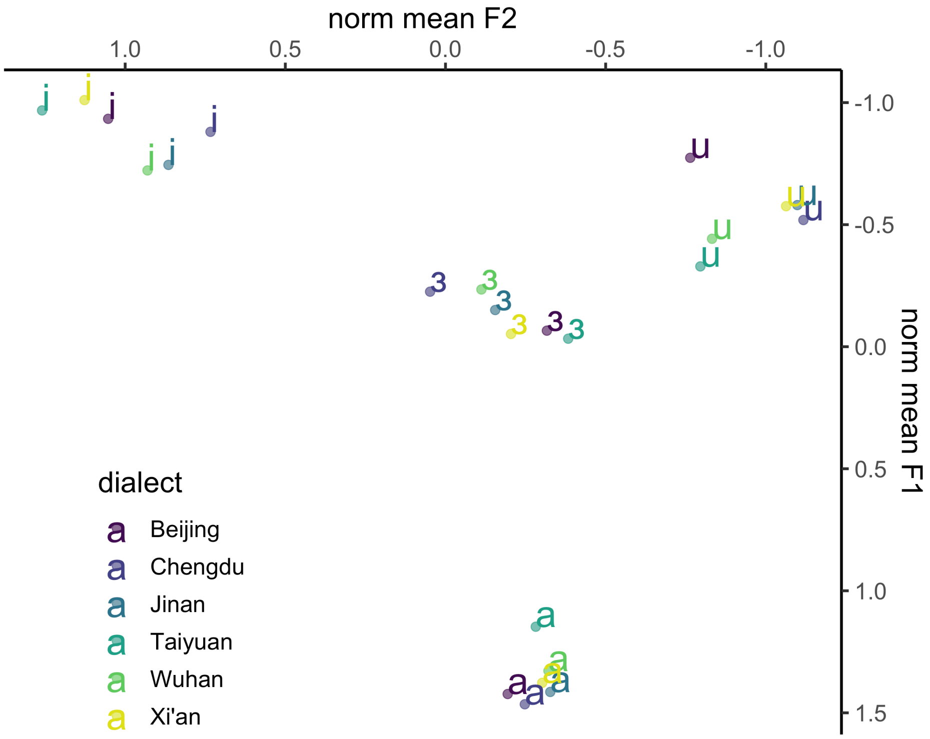


Figure 2.8 The normalised mean F1 and F2 values by vowel and dialect.

The significant effect of vowel for both F1 and F2 measures validated the feasibility of using anchor vowels for approximating the overall vowel space of Mandarin dialects. As shown in Figure 2.8, the six dialects have a similar triangular distribution of [i, a, u] in the acoustic vowel space; the central vowel [ə] appears relatively closer to the back vowel [u] than the front vowel [i]. For dialectal differences, the lack of a significant effect of dialect or its interaction with vowel strongly suggested that Mandarin dialects have a comparable segmental inventory, at least for vowels. Vowel-specific variation within each dialect can be seen in Figure 2.9; the positions of the vowel labels indicated the mean formant values.

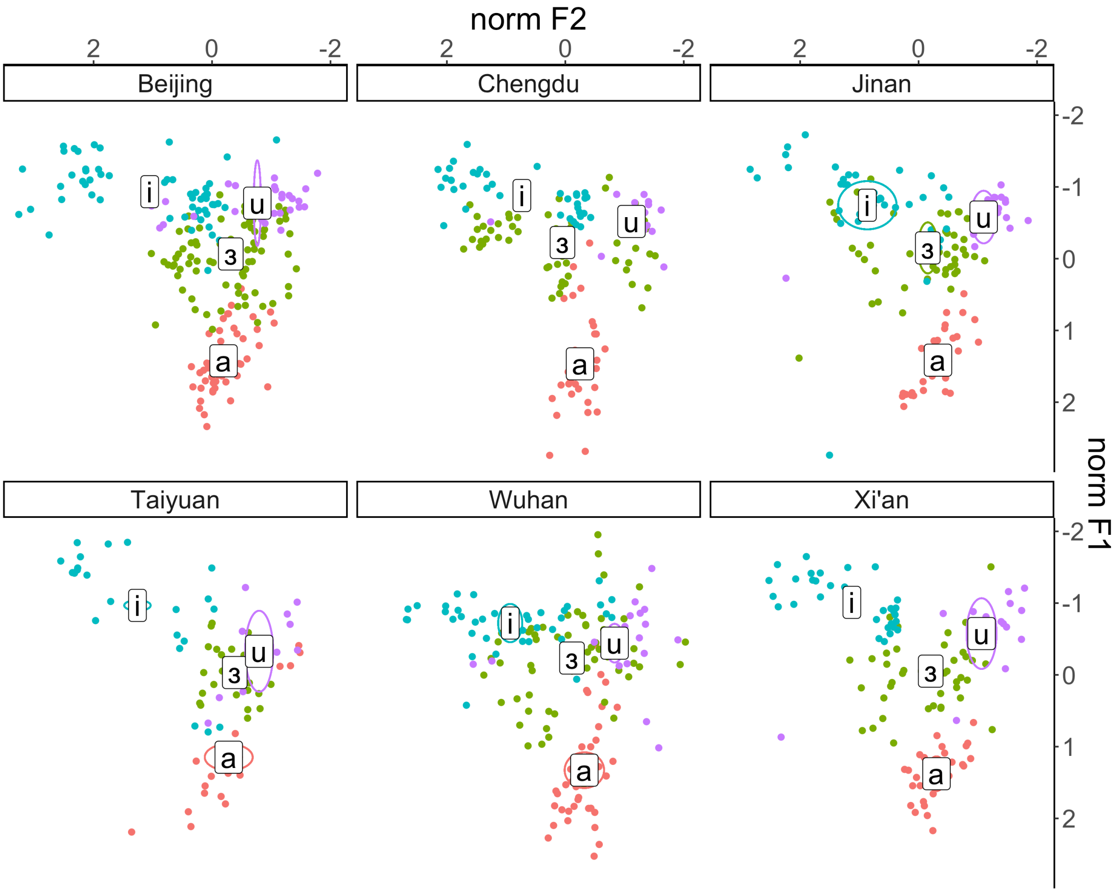


Figure 2.9 Scatterplots of normalised F1 and F2 values by vowel in the six dialects. Ellipses indicate two standard errors from the dialect-vowel-specific mean F1 and F2.

A linear mixed-effects model was implemented to assess vowel duration with fixed effects of vowel, dialect, gender and the interaction between vowel and dialect, and a random intercept for speaker. The coding scheme was the same as the previous models. According to the model, the production of [i] was significantly longer than average in all the six dialects ([i]: $\beta = 0.0509,\ t = 10.579$); the vowel [a] was produced with a significantly shorter duration ([a]: $\beta = -0.0212,\ t = -4.394$); the back vowel [u] did not differ significantly from the mean. With respect to dialectal differences, vowels in Jinan Mandarin had considerably shorter durations (Jinan: $\beta = -0.0807,\ t = -3.471$), while vowels in Wuhan Mandarin were significantly longer (Wuhan: $\beta = 0.0695,\ t = 2.959$) (Figure 2.10). Vowel duration was not significantly modulated by any interaction between vowel and dialect, or gender. Non-normalised mean F1 and F2 values (in Hertz) and durations (in millisecond) are enclosed in Table 2.9.

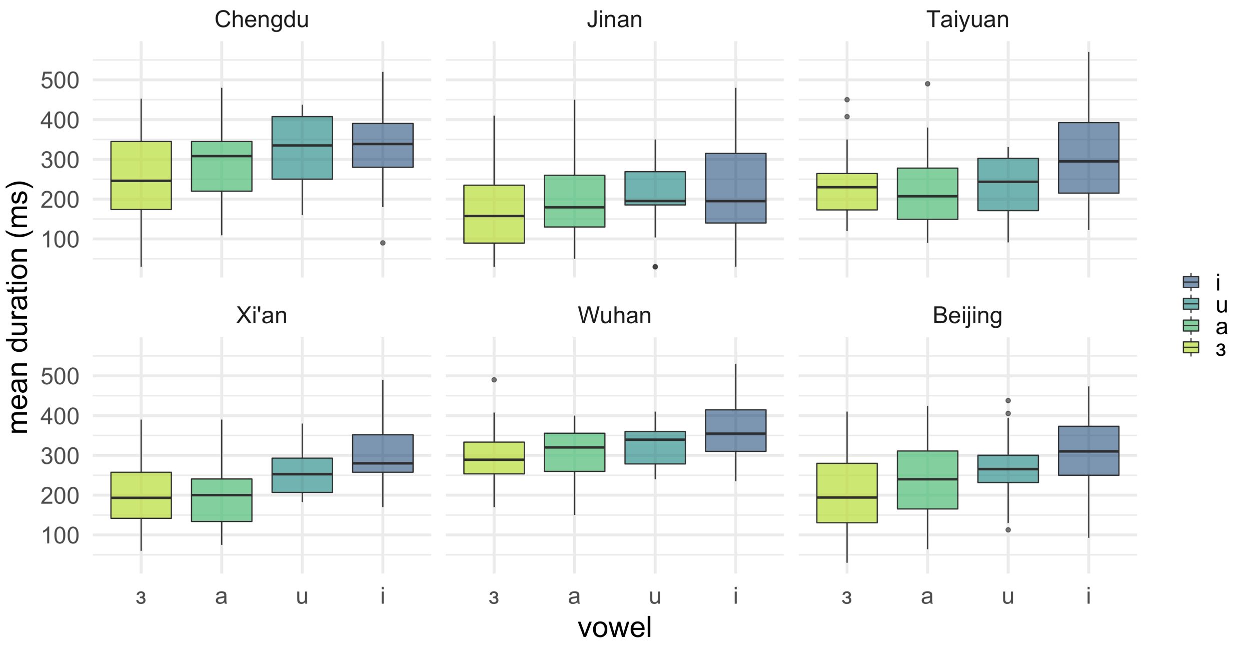


Figure 2.10 Mean vowel durations (ms) across dialects.

Table 2.9 Mean and standard deviation of F1 (Hz), F2 (Hz) and duration (ms) for /i, ə, a, u/ in each of the six dialects.

dialect	vowel	Mean F1	SD F1	Mean F2	SD F2	Mean Dur	SD Dur
Beijing	i	403	76.88	1906	140.72	317	0.02
	u	434	58.38	1032	414.82	271	0.01
	ə	578	54.98	1267	121.47	210	0.02
	a	880	100.22	1321	136.82	241	0.01

Chengdu	i	371	63.95	1936	122.2	341	0.02
	u	453	54.39	954	100.88	321	0.03
	ə	524	67.94	1572	108.08	243	0.02
	a	914	146.41	1417	186.98	287	0.01
Jinan	i	417	79.13	1959	230.44	215	0.03
	u	452	68.78	957	191.87	189	0.02
	ə	547	134.27	1388	173.57	142	0.05
	a	947	166.99	1301	130.41	197	0.05
Taiyuan	i	410	30.98	1944	360.28	312	0.02
	u	573	116.03	980	238.96	239	0.04
	ə	632	104.91	1168	69.21	227	0.02
	a	904	247.96	1237	239.23	221	0.03
Wuhan	i	369	58.31	1938	264.58	367	0.01
	u	435	51.98	1070	138.45	326	0.02
	ə	481	94.94	1427	202.82	286	0.02
	a	840	174.43	1321	125.51	305	0.01
Xi’an	i	382	24.81	2088	254.67	299	0.02
	u	475	39.96	1012	208.13	258	0.04
	ə	593	59.89	1439	98.03	201	0.03
	a	938	182.79	1407	225.81	201	0.02

2.3.2 Phonetic tone inventories

2.3.2.1 Data validation

The production of monosyllabic words in Standard Mandarin from all participants was first analysed to verify the reliability of smartphone recordings for acoustic-phonetic analysis. If the measured pitch contours substantially match the tone system of Standard Mandarin from textbook reference, the collected data should be reliable to represent tone systems of the other Mandarin dialects. Figure 2.11 plots the four lexical tones based on the corpus data, which generally conformed to the relative patterns of Standard Mandarin with Tone 1 as a level tone, Tone 2 a rising tone, Tone 3 a dipping tone, and Tone 4 a falling tone.

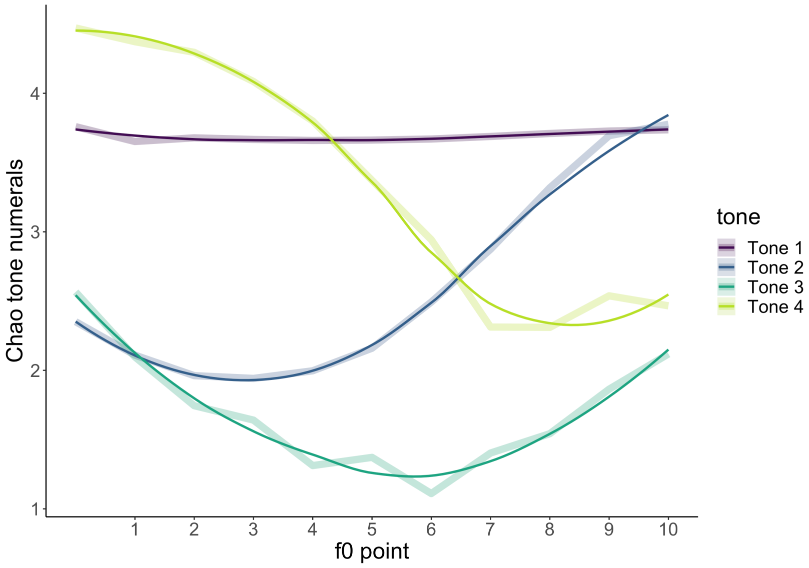


Figure 2.11 Smoothed tone contours of Standard Mandarin based on the corpus data.

2.3.2.2 Dialect-specific tone inventories

Figure 2.12 visualises the pitch contours in semitone by dialect. The six dialects showed distinct phonetic tone inventories. The semitone conversion was applied as the first attempt for visualisation. The semitone-based contours varied across dialects in their relative pitch range; Chengdu and Xi'an Mandarin seemed to have higher overall pitch registers for all tone categories than Beijing, Jinan, Taiyuan and Xi'an. The tone curves also contained a few wiggly portions possibly due to creakily produced sonorant sounds or inaccurate extractions in Praat. To analyse tone contours in a more comparable acoustic range, F0 values were further fitted to the Chao tone scale (for the method, see Section 2.2.5.2) ranging from 0 to 5 to normalise dialect-specific pitch range with a smoothing function on the curves (Figure 2.13).

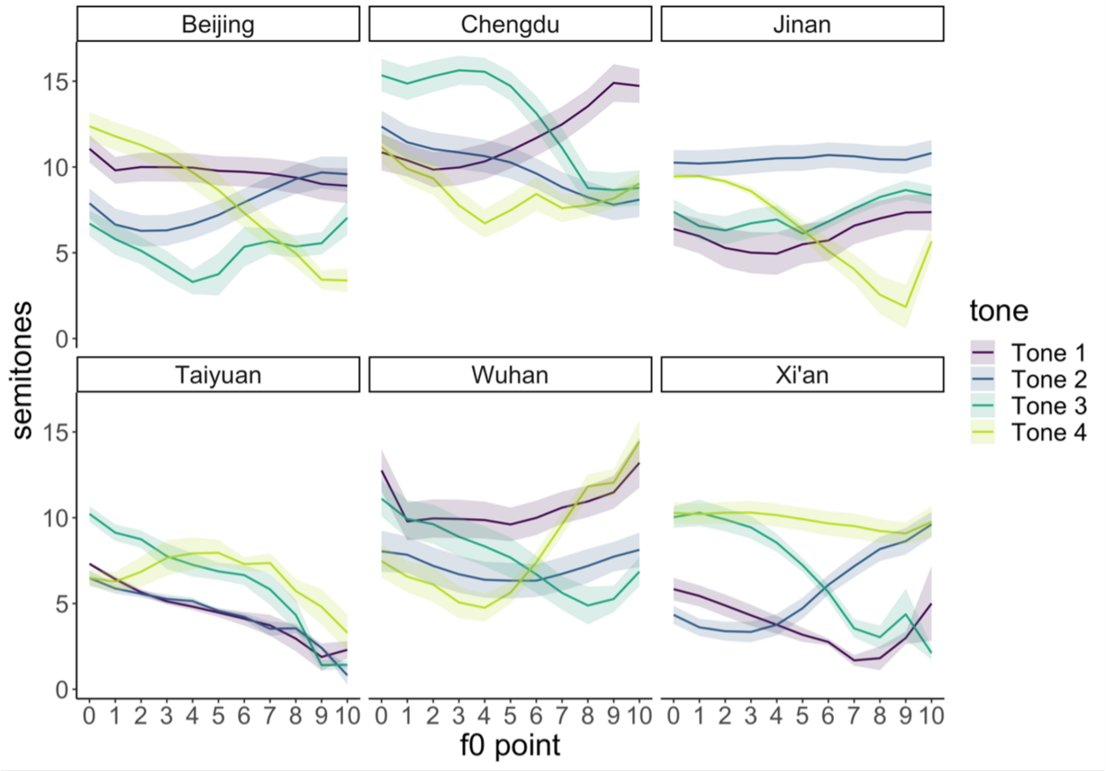


Figure 2.12 Mean F0 (semitone) contours of four lexical tones in six dialects. Ribbons currently reflect ± 0.5 standard error from the mean.

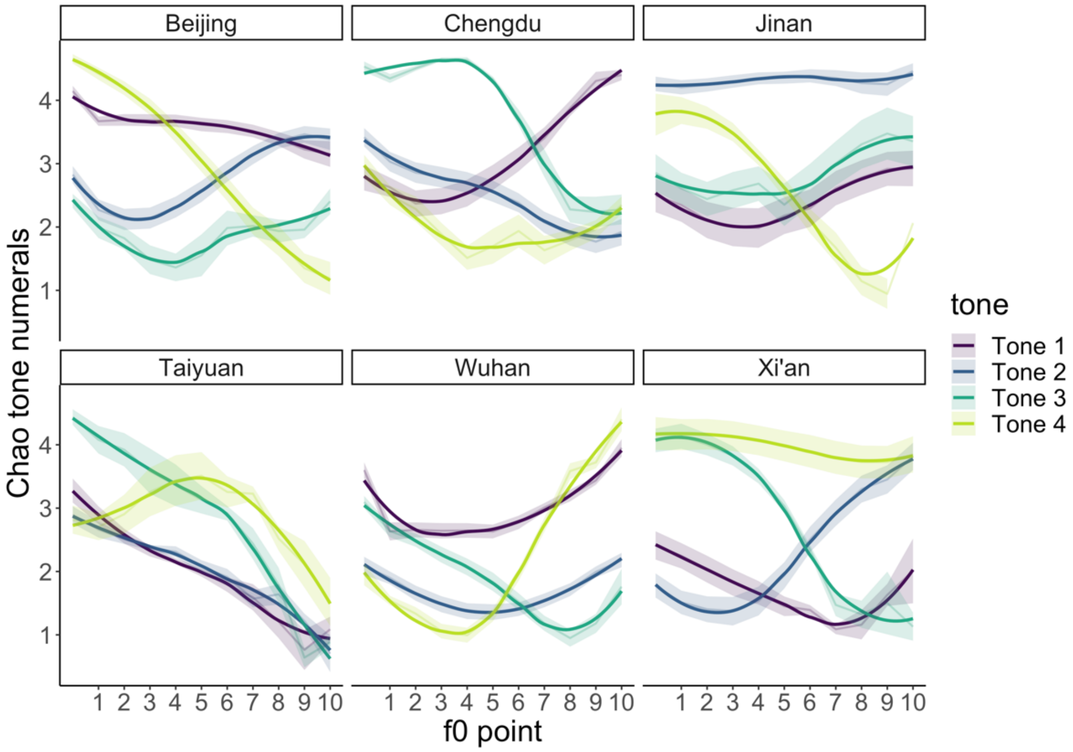


Figure 2.13 Smoothed mean F0 (Chao tone) contours of four lexical tones in six dialects. Ribbons currently reflect ± 0.5 standard errors from the mean.

In Table 2.10, the following is listed for each dialect: 1) previously documented tone contours using the chao tone numerals from *Xiandai Hanyu Fangyan Dacidian* (Modern Chinese Dialect Dictionary, 2012; Ref.1); 2) the researchers' perception of each tone categories; 3) updated Chao tone numerals based on average acoustics from the measurements; 4) descriptions of measured tone contours (Figure 2.10). The contour types were classified as level, rising, falling and dipping. For each contour type, F0 onset was indicated as either low or high based on the measured acoustics, or unspecified if this contour type only shows in one tone category or does not differ in the onset value between multiple tone categories. For example, Tone 2 and Tone 3 in Chengdu Mandarin both had a falling contour, but the onset pitch value of Tone 2 was considerably lower than Tone 3; Tone 2 was therefore indicated as a low-falling tone, and Tone 3, a high-falling tone. There were a few cases (marked by an asterisk) where the general contour patterns from data did not match previous records. This might be attributed to phonetic variation over time or due to the relatively small sample size.

Table 2.10 Comparison of tone systems from different sources. Asterisks mark the tones that show completely different contour patterns between measured and recorded tones.

	source	Tone 1	Tone 2	Tone 3	Tone 4
Beijing	Ref. 1	55	35	214	51
	perception	level	rising	dipping	falling
	measured	44/43	24	212	51
	production	level	rising	dipping	falling
Chengdu	Ref. 1	55	21	53	213
	perception	rising	low-falling	high-falling	dipping
	measured	*35	32	52	323
	production	rising	low-falling	high-falling	dipping
Jinan	Ref. 1	213	42	55	21
	perception	dipping	level	rising	falling
	measured	323	*55	*34	41
	production	dipping	level	rising	falling
Taiyuan	Ref. 1	11		53	45
	perception	low-falling		high-falling	rising
	measured	*31		51	34(2)
	production	low-falling		high-falling	rising(-falling)
Wuhan	Ref. 1	55	213	42	35
	perception	rising/level	dipping	falling	rising
	measured	(4)35	212	31	(2)15
	production	rising/level	dipping	falling	rising
Xi’an	Ref. 1	21	24	53	44
	perception	low-falling	rising	high-falling	level
	measured	31	24	51	55
	production	low-falling	rising	high-falling	level

For each dialect specifically, the measured pitch contours of Beijing Mandarin generally matched previous descriptions and the researcher's perception. Four distinct types of contour patterns were all present in this dialect, although Tone 2 in production might not reach the highest F0 as illustrated by [35], but more of [23] or [24] possibly due to tone declination over time of utterance production. Such declination also occurred on Tone 1—Tone 1 stayed as a high-level tone with a slight gradual decline towards the end of the word production. Tone 3 showed a clear dipping contour but with the onset and the ending point on relatively the same pitch level, rather than [214] as previously known. Tone 3 commonly co-occurred with creaky voice, which might be a secondary cue to Tone 3 especially when Tone 3 is realised in the lower range of F0 (Zhang & Kirby, 2020).

For Chengdu Mandarin, the measured tone contours matched the perceived patterns, and had only one different pattern from the previous descriptions, which was for Tone 1. Tone 1 was formerly recorded as a level tone [55], but both the acoustics and the perception indicated a rising contour for Tone 1. Moreover, a distinct onset difference was found between Tone 2 and Tone 3; the pitch onset of Tone 2 ([32]) was considerably lower than that of Tone 3 ([52]). The falling pattern of Tone 3 was particularly interesting because the curve did not drop linearly from a higher to a lower pitch, but plateaued over the initial period at a high pitch level and then dropped gradually towards the lower pitch. For Tone 4, the pitch contour was not as curvy or dipping as might be suggested by [213] and was instead measured as a [323] pattern.

Jinan Mandarin showed differentiating patterns of the tone contours between previous records and the measured data except for Tone 1. Tone 2 was realised as a high-level tone instead of a falling [42]. Tone 3 was measured and perceived as slightly rising, closer to a [34] pattern instead of [55]. Tone 4 from the production data showed the same falling contour as previously recorded, but fell across a much larger range of F0 rather than [21]. A particular observation of Jinan tone system is that it makes use of a similar set of contour types as Standard Mandarin, with only the mappings between phonological categories and phonetic realisation were re-arranged. Standard Mandarin Tone 1 resembles Jinan Tone 2, and the same happens between Standard Tone 2 and Jinan Tone 2, and between Standard Tone 3 and Jinan Tone 1. Tone 4 in the two dialects has a similar falling contour. Compared to the other dialects, lexical tones in Jinan Mandarin seemed to be phonetically more similar to those in Standard Mandarin. For native speakers of Standard Mandarin, if presented with Jinan tones in isolation, it is highly likely that the listeners would identify those tones as from their native Standard tone system.

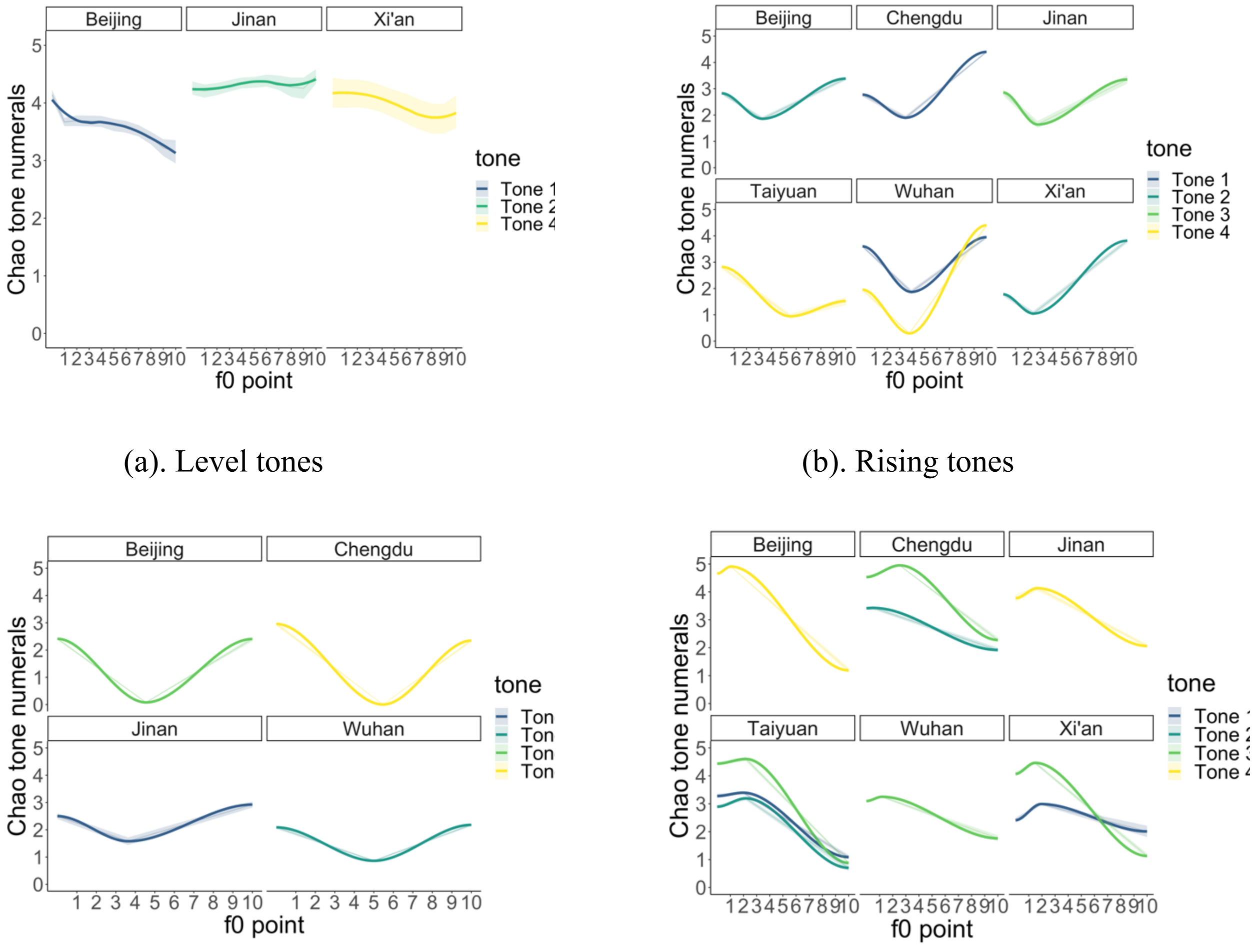
Taiyuan Mandarin stood out from the other dialects with a three-tone system, where Tone 1 and Tone 2 merged as a low falling contour ([31]). Taiyuan Tone 3 showed a high-falling pattern with the falling tail pointing to the lowest pitch range. Tone 4 was perceived as a rising tone despite the falling tail towards the end of the word production. The rising-falling pattern was not typically used for lexical tone contrast in Mandarin dialects because the falling contour coincides with the natural decline of pitch towards the end of a phrase, and tone systems tend to avoid unnecessarily complex contour types that might create ambiguity when interacting with intonational patterns.

For Wuhan Mandarin, Tone 1 was perceived as either a level or a rising tone depending on different lexical words; the rising part appeared quite late towards the end of the word. Tone 2 was a dipping tone closer to [212] rather than [213]. Tone 3 was consistent across all the sources. Tone 4 was perceived as a rising tone, but the produced pitch contour contained an initial falling part with a seemingly dipping pattern. Alternatively, this could be interpreted that a dipping contrast is present in Wuhan Mandarin—Tone 2 is central-dipping with the turning point of F0 around the mid of the curve and Tone 4 is front-dipping with the turning point closer to the onset (Zhu, Yi, Zhang, Nguyễn, 2019).

The measured tone system of Xi'an Mandarin was consistent between production and perception. Tone 1 and Tone 3 contrasted in the onset pitch level, and both showed a falling pattern despite a brief rising part towards the end of the curve. Tone 2 was a rising tone and Tone 4 was a high-level tone. These findings aligned with previous descriptions, including the more recent phonetic study of Xi'an Mandarin tones in Liu et al., 2020.

2.3.2.3 Phonetic similarity of tones with the same contour type

Phonetic tones in the six dialects demonstrated four contour patterns—level, rising, dipping and falling. Tones of the same contour type varied in terms of the pitch onset, the steepness of the slope, the turning point and ΔF0 (Figure 2.14). Mixed-effects second-order polynomial models were implemented for each contour type to predict the pitch trajectory from F0 point and dialect. Each model included the fixed effects of dialect, F0 point and the interaction between dialect and F0 point, and a random intercept for speaker. Sum coding was used to compare individual dialect level to the grand mean. The second-order orthogonal polynomial models use a multilevel regression technique designed for analysing time course data (Mirman, 2017; Rattanasone et al., 2018; Tang et al., 2019). According to Mirman (2017), the polynomial function generates three coefficients: the intercept term (i.e., pitch onset), the first-order linear term (i.e., pitch slope), and the second-order quadratic term (i.e., pitch curvature). More specifically, a positive linear coefficient indicates a rising pitch contour, whereas a negative linear coefficient indicates a falling contour; a larger absolute value of the linear coefficient represents a steeper slope. A positive quadratic coefficient indicates a concave contour and a negative one indicates a convex contour; a larger absolute value of the quadratic coefficient indicates curvier contours.



(a). Level tones

(b). Rising tones

(c). Dipping tones

(d). Falling tones

Figure 2.14 Smoothed Chao tone contours by the four contour patterns (a-d) in the six dialects.

For the level contours, the model detected a significant intercept ($\beta = 3.94$, $t = 6.75$), a significant linear slope ($\beta = -1.92, t = -4.39$) and its interaction with Jinan Mandarin ($\beta = 2.38$, $t = 3.85$). The significant positive intercept indicated a consistent high pitch onset of level tones across the three dialects. The significant linear trend in the absence of a significant quadratic effect suggested a linear contour for the level tones as expected. The linear slope was also significantly modulated by dialect; the level tone of Jinan Mandarin differed from the others with a significant rising trend.

The results for the rising tones showed significant effects for both the linear ($\beta = 9.34$, $t = 11.07$) and the quadratic trends ($\beta = 3.65$, $t = 4.32$), indicating a convex curve for the rising tones, which is consistent with the previous findings that rising tones in Mandarin typically involve a brief falling part from the pitch onset. Significant interactions with dialect were found for both the linear and quadratic trends. For the linear interactions, the positive effect on the linear interactions with Chengdu, Wuhan and Xi'an Mandarin suggested steeper rising parts for the rising tones in these three dialects relative to the average. For the quadratic interactions, the negative effect on the quadratic slope for Taiyuan Mandarin ($\beta = -1.308e + 01$, $t = -5.65$) indicated a flatter curve for Taiyuan Tone 4, whereas the positive effect for Wuhan ($\beta = 6.536e + 00$, $t = 4.50$) suggested more curvy rising contours.

For the dipping tones, significant effects emerged for the intercept ($\beta = 1.99$, $t = 11.78$) and the quadratic slope with a positive estimate on the quadratic term ($\beta = 5.23$, $t = 6.69$), which indicated a convex pattern for all the dipping tones. Moreover, although significant linear interactions with dialect were found according to the model, the estimates for the linear slope were in the expected direction. Nevertheless, based on the dipping contours in Figure 2.11 (c), the dipping tones in Beijing and Chengdu Mandarin were curvier than those in Jinan and Wuhan Mandarin.

Since multiple falling contours were identified in the six dialects, the polynomial model for the falling pattern added an additional factor for dialect-specific tone categories and the full interactions with the linear and quadratic terms and dialect. Significant effects were found for the linear slope and its interaction with tone and dialect. The negative estimate on the linear trend suggested a significant falling pattern across tones and dialects. For Chengdu, Taiyuan, and Xi'an specifically, Tone 3 was produced with a steeper drop in pitch indicated by the significant linear trends with larger absolute values for Tone 3.

2.3.2.4 Edit distance of tone systems across dialects

With the updated tone systems (Table 2.11), tone-specific edit distance was calculated between Standard Mandarin and each Mandarin dialect for each tone category. Both the original and the revised calculating methods were applied (Table 2.12). Tone-specific edit distance was then averaged by the number of tone categories to compute the dialect-specific distance relative to the tone system of Standard Mandarin. A lower value of the average distance indicates greater similarity of the dialect-specific tone system to Standard Mandarin.

Table 2.11 Updated Chao tone numerals based on measured acoustics (see also in Table 2.10).

	Tone 1	Tone 2	Tone 3	Tone 4
Standard Mandarin	44	24	212	52
Beijing Mandarin	44	24	212	51
Chengdu Mandarin	35	32	52	323
Jinan Mandarin	323	55	34	41
Taiyuan Mandarin	31		51	34(2)
Wuhan Mandarin	(4)35	212	31	215
Xi’an Mandarin	31	24	51	55

Table 2.12 Average tone edit distance of the six dialects relative to Standard Mandarin.

(a) The original method

	Beijing	Chengdu	Jinan	Taiyuan	Wuhan	Xi’an
Standard Mandarin	0.208	0.915	0.833	0.938	0.688	0.563

(b) The revised method

	Beijing	Chengdu	Jinan	Taiyuan	Wuhan	Xi’an
Standard Mandarin	0.208	1.458	1.667	1.833	1.583	1.208

According to Table 2.12 (a) and (b), Beijing Mandarin had the most similar tone system to Standard Mandarin as expected given that Standard Mandarin was developed based on the sound inventories of Beijing Mandarin. Xi’an Mandarin had a relatively more similar tone system to Standard Mandarin than the other dialects excluding Beijing Mandarin. It seemed that the tone system of Taiyuan Mandarin deviated most from Standard Mandarin compared to the other dialects. For the tone systems of Chengdu and Jinan Mandarin, their relative similarity to Standard Mandarin varied between the two proposed methods.

2.4 Conclusion

This chapter investigated the vowel spaces, vowel duration and tone inventories across six Mandarin dialects. Speech data were collected over controlled productions of monosyllabic words by native speakers of the six dialects. Relative positions of monophthongs /i, ə, a, u/ were compared within the F1×F2 acoustic spaces across dialects. It was established that Mandarin dialects share a similar vowel inventory with monophthongs positioned at approximately the same areas within the vowel spaces across dialects. Among the measured vowels, central vowels including [ə] and [a] were most stable across dialects, whereas the relative positions between the front vowel [i] and the back vowel [u] slightly varied among dialects. It seems that as the front vowel was further estranged from the central ones, the back vowel tended to cluster closer to the central vowels.

With regards to the tone systems, substantial differences were observed across dialects. The measured pitch contours generally conformed to the researchers' perception and were mostly consistent with previous records despite a few discrepancies between observed contours and documented ones. F0 values were taken from the same lexical items produced by speakers of the six dialects. Plotted pitch contours demonstrated a disparate phonetic tone inventories of the six dialects.

Quantification of phonetic similarity was applied for both contour-based and dialect-based analyses. Considerable phonetic variation was found across dialects and across tone categories of the same contour. Although F0 contours were used as the primary acoustic cue in lexical tone production, tone systems with more than one category of the same contour type exhibit finer F0 characteristics, such as F0 onset and the steepness or the curvature of the slopes. Direct comparisons of the dialectal tone systems with Standard Mandarin using edit distance to some extent revealed the between-dialect distance of the tone systems, but the results were far from being sufficient to reliably predict which dialect has a more similar tone system to Standard Mandarin. However, we should be able to conclude based on the measured contours that tone systems of certain Mandarin dialects, such as Jinan and Xi'an Mandarin, are phonetically more similar to that of Standard Mandarin, while dialects like Chengdu and Taiyuan Mandarin use different phonetic contours for their tone systems compared to Standard Mandarin.

2.5 Authorship and publication status

This chapter contains content published in the peer-reviewed paper "The ManDi Corpus: A Spoken Corpus of Mandarin Regional Dialects" in the Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), co-authored by Liang Zhao and Eleanor Chodroff. The relative contribution is as follows:

Liang Zhao: Conceptualisation, methodology, investigation, resources, data curation, formal analysis, visualisation, writing—original draft, writing—review & editing; Eleanor Chodroff: methodology, formal analysis, visualisation, writing—review & editing, supervision, funding acquisition.

Chapter 3 Perception of lexical tone variation with incidental exposure

3.1 Introduction

As discussed in Chapter 2, Mandarin dialects are mutually intelligible language varieties with comparable segmental and highly disparate tonal realisations. For native speakers of Standard Mandarin or any Mandarin dialect, understanding speech in an unfamiliar Mandarin dialect primarily involves dealing with a familiar segmental system, but a rather unfamiliar tone system. We might ask: How are the unfamiliar tones processed? Which mechanisms are being used? Can listeners adapt to the new tone system with certain amount of input? To answer these questions, Chapter 3 investigates the perceptual mechanisms involved in processing unfamiliar phonetic tones and the potential factors that might affect the perception outcome. The following introduction sections bring back the two higher-level perceptual mechanisms reviewed in Chapter 1 (Section 1.3) and highlight the influential factors in unfamiliar speech adaptation given previous research on perceptual learning/adaptation (Section 1.4). Sections 3.2 and 3.3 present two perception experiments using the sentence semantic-plausibility judgment task with Standard Mandarin and Chengdu Mandarin stimuli. The first study (Section 3.2) revealed listeners' rapid adaptation to unfamiliar phonetic tones and the integrated use of top-down and bottom-up information in processing familiar and unfamiliar tone systems. The second study (Section 3.3) followed up on the results of the first experiment and manipulated the sentential stimuli to further examine the effects of quality and quantity of exposure on the adaptation outcome. Section 3.4 provides a concise conclusion of the two experiments and major findings. The authorship and publication status of the content can be found at the end of each research chapter.

3.1.1 Top-down and bottom-up processing of speech

In addressing how the speech signal is processed in decoding the intended utterance, several prominent theoretical frameworks have identified two high-level mechanisms: top-down processing and bottom-up processing. Bottom-up processing refers to the processing of acoustic properties from the immediate incoming signal; top-down mechanism refers to processing based on prior knowledge, either linguistic knowledge or world knowledge. Early models of speech perception, such as the Cohort model (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987), Direct Perception (Gibson, 1954) and Direct Realism (Fowler, 1986) have often assumed a privileged role of bottom-up information in the perceptual system. However, subsequent theories have taken the influence of top-down information into consideration, e.g., TRACE (McClelland & Elman, 1986) and Acoustic Landmarks and Distinctive Features (Stevens, 2002). Others have eschewed any role for top-down processing, e.g., Shortlist (Norris, 1994) and Merge (Norris, McQueen & Cutler, 2000). Differing views on the inclusion of top-down process was summarised in Table 1.1. Current literature tends to incorporate both top-down and bottom-up mechanisms to reflect the constructive nature of human speech perception (Section 1.3), but the relative weighting and integration of these sources of information remains unclear.

Moreover, previous work on speech perception has been historically segment-oriented and concentrated on non-tonal languages, leaving the mechanisms for tonal speech perception relatively under-investigated. For tonal languages such as Mandarin dialects, lexical tones are crucial to differentiate the meanings of lexical items. The fact that tonal information is heard and perceived alongside of, and not independent from, segmental information has given rise to research on the relative weighting of lexical tone and segmental information for lexical access. Some have argued that segmental information is more salient than tonal information in sub-lexical processing as tonal information is accessed later or with lower accuracy than segments (Taft & Chen, 1992; Cutler & Chen, 1997; Ye & Connine, 1999; Wiener & Turnbull, 2016). Specifically, Taft and Chen (1992) reported substantial difficulty in discriminating homophones presented in written and spoken Chinese characters particularly when the two characters were paired with identical phonemes, but different tones. Longer latency and lower accuracy were also found for tonal contrasts relative to segmental contrasts in the word-nonword decision, same-different character judgment (Cutler & Chen, 1997), word monitoring with no or neutral context (Ye & Connine, 1999), and word reconstruction tasks (Wiener & Turnbull, 2016). The tonal disadvantage was particularly noticeable when the input information is lacking in lexical context (see also in Section 1.2.1).

Others dispute this seemingly inferior status of tones in lexical access, and contend that lexical tones could have an equal or even greater contribution to lexical access relative to segments given appropriate top-down feedback (Liu & Samuel, 2007; Malins & Joanisse, 2010). Lexical judgments on disyllabic words and idioms were equally accurate for segmental and tonal manipulations in Liu and Samuel's study (2007). Results from an eye-tracking study, in which participants matched a spoken word to an array of pictures, also indicated a comparable contribution of segmental and tonal information in lexical access (Malins & Joanisse, 2010). More recently, the Reverse Accessing Model (RAM) reported a distinct advantage for segments over lexical tones in terms of information accessing, suggesting that tone information is accessed only *if necessary* (Gao et al., 2019; see also in Section 1.2.1).

3.1.2 The effect of quality and quantity of exposure

Perceptual adaptation to unfamiliar speech requires adequate exposure to the target speech especially when the initial input is received later in adulthood. What makes the exposure "*adequate*" has become a central question to research on speech adaptation. Previous findings suggest that the adaptation outcome may be modulated by both quality and quantity of the spoken stimuli.

Quality refers to the type, structure, and source of information in the experimental exposure. For example, lexical information has been considered beneficial for adaptation to a novel sound contrast (Norris et al., 2003; Hayes-Harb, 2007). Hayes-Harb (2007) tested English speakers' discrimination of [g]- and [k]-like novel sounds after auditory training either with minimal pairs of the sounds, or with members of a [g]-[k] continuum without lexical meaning in a bimodal distribution that favoured tokens towards the endpoints of the continuum. The results showed that adaptation to the novel contrast occurred with statistical learning alone, but discrimination was significantly enhanced when a lexical contrast was present. That said, listeners are able to rapidly discriminate between previously unheard linguistic contrasts: listeners are able to generalise heard patterns to new segments (Maye et al., 2008) and words unheard in the training (McQueen, Cutler & Norris, 2006). Nevertheless, it is likely that the adaptation process could be facilitated if an explicit lexical contrast, such as a minimal pair, was heard in the exposure.

Listeners also frequently rely on explicit training to gather ample information for adaptation. If more information is available in the test stimuli, such as in sentences and passages, adaptation may happen without explicit training. Clarke & Garrett (2004) reported listeners' rapid adaptation to Spanish- and Chinese-accented speech with one minute of incidental exposure to the sentence stimuli. In addition, exposure with about sixteen sentences was found sufficient to initiate adaptation to a foreign-accented talker (Bradlow & Bent, 2008). Crucially, however, listeners significantly improved over the course of the experiment. Even though adaptation can be reasonably successful in a short period, increased exposure may help listeners to generalise heard patterns to novel sounds or speakers.

The introduction chapter of the thesis, particularly Section 1.4, has provided more detail on these studies on adaptation to novel speech or accent. To conclude here, previous findings have suggested that enhancement in the quality and quantity of the exposure stimuli can help improve perceptual adaptation, but it is unclear to what extent the stimuli should be manipulated and how much input is sufficient to observe successful adaptation. The specific research questions will be explained in the following Sections 3.2 and 3.3 for each experiment respectively.

3.2 Experiment 1: Top-down and bottom-up processing of familiar and unfamiliar tone systems

While researchers generally agree that both top-down and bottom-up information are used in speech processing, there is little consensus on the relative weighting of these sources of information and how they interact. To what extent do speaker expectations guide (or indeed, override) attendance to bottom-up segmental and pitch information? To test this, we manipulated the reliability of tonal information in high and low surprisal (lexical expectedness) sentences using natural regional variation among Mandarin dialects in a sentence semantic-plausibility judgment task. Mandarin dialects provide a natural testbed for research on tonal speech perception due to their comparable segmental inventories, but distinct tone systems. We focused on Standard Mandarin and Chengdu Mandarin. Both dialects have a four-tone system and the same mapping between phonological tone category and lexical category. They differ, however, in the phonetic implementation of each phonological tone category. According to the measured Chao tone numerals (Table 2.11), Standard Mandarin has Tone 1 ([44]), Tone 2 [24], Tone 3 [212], and Tone 4 [52], whereas Chengdu Mandarin has Tone 1 [35], Tone 2 [32], Tone 3 [52] and Tone 4 [323] (see also in Figure 3.1).

For the familiar tone system (Standard Mandarin), we expect speakers to use both top-down and bottom-up information for lexical access, as the sentential context and tonal representations are both reliable cues for native listeners. For the unfamiliar tone system (Chengdu Mandarin), we expect the dominance of top-down information from sentential context, and little or no use of lexical tone due to the unfamiliarity of the tone system.

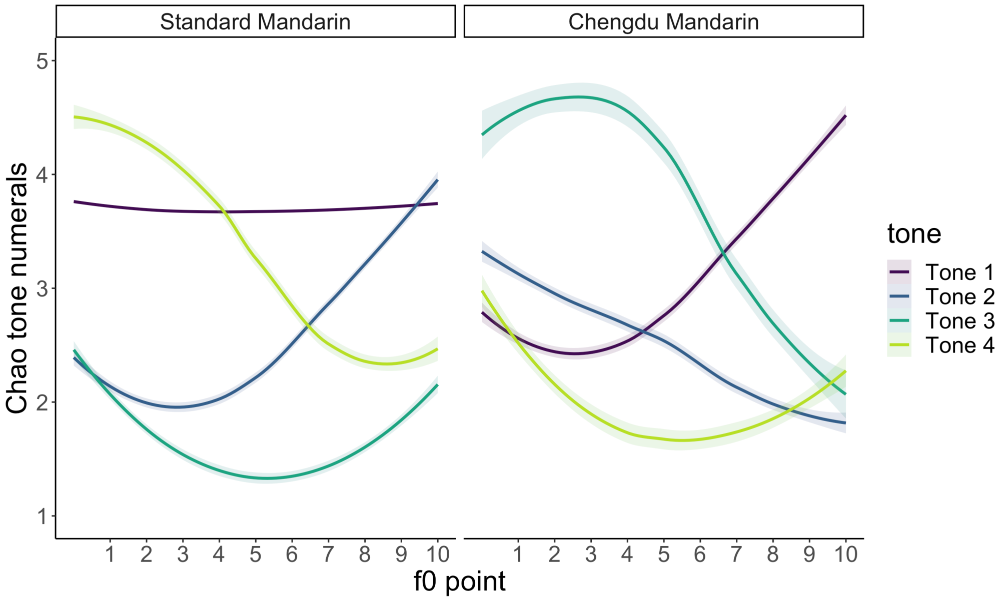


Figure 3.1 Smoothed lexical tone contours of Standard Mandarin and Chengdu Mandarin converted to Chao Tone numerals (Zhao & Chodroff, 2022). Ribbons reflect ± 0.2 standard error from the mean.

3.2.1 Methods

A 2×2 factorial design was used to assess the effects of sentence semantic plausibility (high surprisal vs. low surprisal) and dialect familiarity (native Standard Mandarin vs. non-native Chengdu Mandarin) on the accuracy of semantic plausibility judgments and response times.

3.2.1.1 Participants

Twenty-one native speakers of Standard Mandarin who reported little or no knowledge of Chengdu Mandarin participated in the experiment. No participant reported hearing or reading impairments.

3.2.1.2 Materials

Twenty-four sentences were created manipulating Mandarin Dialect in a between-item design (12 Standard Mandarin and 12 Chengdu Mandarin sentences). Within these sentences, the lexical tone of a critical word was manipulated resulting in either a semantically plausible (low surprisal) or a semantically implausible (high surprisal) sentence. Participants heard different sets of sentence items in Standard Mandarin (native dialect) and Chengdu Mandarin (non-native dialect) trials. Half the critical words were sentence-medial and half were sentence-final. Tone combinations were counterbalanced across items.

Table 3.1 gives an example pair of high and low surprisal sentences, presented in simplified Chinese characters and *Pinyin* orthography with its tone category—tone 1, tone 2, tone 3, and tone 4. Note that participants only heard the auditory version of the sentence. The phonetic tone realisations of these words in Chengdu Mandarin are considered unknown or unfamiliar to speakers of Standard Mandarin (see Figure 3.1). Surprisal was manipulated by altering the tone of a critical word in which the segments were rendered intact, but were paired with different tones. In the example here, /fei1/ (plausible: *"There is an eagle in the sky flying"*) contrasted with /fei2/ (implausible: *"There is an eagle in the sky gaining weight"*). Participants heard both renditions of each sentence for a total of 48 trials.

Table 3.1 An example sentence item across surprisal conditions.

low-surprisal sentence	a) 有 一只 鹰 在 天上 飞 You3 yi4 zhi1 ying1 zai4 tian1 shang4 fei1 There is an eagle in the sky flying “There is an eagle flying in the sky”
high-surprisal sentence	b)* 有 一只 鹰 在 天上 肥* You3 yi4 zhi1 ying1 zai4 tian1 shang4 fei2* There is an eagle in the sky gaining weight* “There is an eagle gaining weight in the sky”

The Standard Mandarin stimuli were produced by a female native speaker of Standard Mandarin (aged 26); Chengdu Mandarin stimuli were produced by a male native speaker of Chengdu Mandarin (aged 29). A 10-ms silence was inserted at the beginning of each sentence, and the audio file was scaled to an intensity of 70 dB.

3.2.1.3 Procedure

The experiment was run online using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Participants were asked to complete the sentential semantic plausibility judgment task on a device with internet access in a quiet environment, and with headphones if possible. They were first briefed on the purpose and content of the experiment. The participants were made aware that they would be listening to sentences spoken in either Standard Mandarin or another Mandarin dialect. Then they were presented with a test audio and adjusted the volume of the sound output to a comfortable level.

In the practice phase, participants listened to two example pairs of high-surprisal and low-surprisal sentences in Standard Mandarin, answered the question “*Does this sentence make sense*”, and then received feedback regarding their answer in the form of a written version of the sentence. Specifically, participants were instructed to click the “play” button to start the audio and then click on the “yes” or “no” button on the screen to answer the question. The correct answer and the sentence in standard simplified Chinese characters were then presented. The “yes” and “no” buttons were presented closely adjacent to each other at the center of the screen with the “yes” button on the left side and the “no” button on the right side.

In the test phase, the presentation of trials was fully randomised. The procedure was identical to the familiarisation stage except that no feedback was provided.

3.2.1.4 Data analysis

Accuracy and response time from the test phase were analysed as dependent variables across manipulations of dialect (Standard Mandarin vs. Chengdu Mandarin) and surprisal (high-surprisal vs. low-surprisal).

“Yes” responses to low-surprisal (i.e., plausible) sentences and “No” responses to high-surprisal (i.e., implausible) sentences were coded as “correct” responses. Response time was calculated as the interval between the end of the audio file and the click registering a judgment. Five trials were excluded from the analysis due to missing or negative response times, likely due to internet connectivity issues. One sentence pair in each dialect was also omitted due to experiment error. This left a total of 919 trials (range of 43–44 trials per participant) for analysis.

Accuracy was modelled with a Bayesian logistic mixed-effects regression, and response time with a Bayesian log-normal mixed-effects regression, both with weakly informative priors[5] (Bürkner, 2018). Each model included fixed effects of surprisal, dialect,

[5] Weakly informative priors are used to trim down the extreme values and not to bias the tested parameters. Prior sensitivity analysis was done to set appropriate priors for the Bayesian models in Chapters 3 and 4. The current priors were mildly informative for the accuracy models and inclined towards principled priors for the RT models, i.e. positive RT values and normal distribution for the intercept based on mean of the data. See more in Nicenboim, Schad & Vasishth’s (2024) work.

trial number, and the full set of interactions. The random effect structure for participant included an intercept and slopes for surprisal, dialect, trial number and the interaction between surprisal and dialect, and for sentence frame, an intercept and random slope for dialect. Priors for main effects and interactions were Normal distributions centred on 0 with a standard deviation of 20 for the accuracy model ($\mathcal{N}(0, 20)$) and Normal distributions centred on 0 with a standard deviation of 1 for the response time model ($\mathcal{N}(0, 1)$). The prior for the intercept was $\mathcal{N}(0, 20)$ for accuracy and $\mathcal{N}(7, 1)$ for response time. The model was run for 2000 iterations with a burn-in period of 1000 iterations. Surprisal and dialect were sum-coded (*surprisal:* high-surprisal = 1, low-surprisal = -1; *dialect:* Chengdu = 1, Standard Mandarin = -1), and trial number was centered on the mean. If the 95% credible interval for an estimated effect excluded 0 (i.e., no effect), then it was deemed to be *credible* in its direction of influence on the respective dependent variable.

3.2.2 Results

3.2.2.1 Accuracy (Experiment 1)

As shown in Figure 3.2, accuracy was near ceiling for both surprisal conditions in Standard Mandarin (high: 98%, low: 92%), but differed considerably by surprisal in Chengdu Mandarin (high: 20%, low: 94%). For the familiar speech (Standard Mandarin), the overall high accuracy suggests that participants understood the task in general, validating the plausibility of the surprisal manipulation in the experiment.

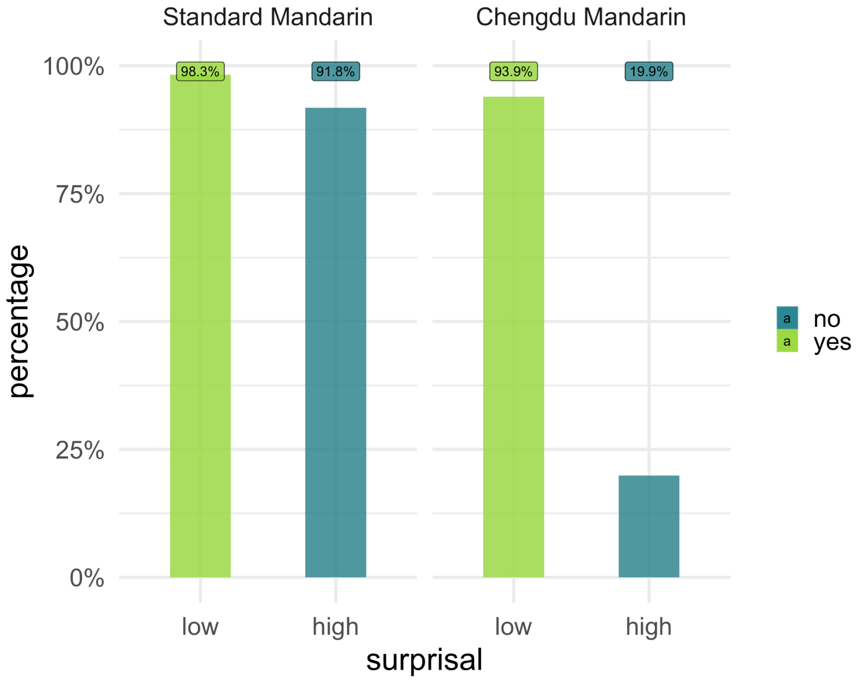


Figure 3.2 Percentage of “correct” responses across dialect and surprisal conditions. “Yes” (plausible; lime green) is treated as correct for low-surprisal conditions, and “no” (not plausible; blue green) for high-surprisal conditions.

Correspondingly, the model revealed credible main effects of surprisal, dialect and the interaction between surprisal and dialect on accuracy. Specifically, accuracy was higher in the low-surprisal condition than in the high-surprisal condition (*surprisal:* β = −2.21, 95% CI = [−3.43, −1.34]). In addition, accuracy was higher for sentences spoken in Standard Mandarin than in the Chengdu dialect (*dialect:* β = −1.96, 95% CI = [−2.86, −1.22]). A credible interaction was also observed between surprisal and dialect, indicating an even lower accuracy for sentences spoken in the Chengdu dialect in the high-surprisal condition (*surprisal x dialect:* β = −1.00, 95% CI = [−1.88, −0.09]). Trial number and its interactions with surprisal and dialect were not reliable in the direction of their effects, indicating that accuracy did not reliably improve in any condition across the course of the experiment (*trial:* β = 0.03, 95% CI = [−0.01, 0.07], *trial x surprisal:* β = 0.02, 95% CI = [−0.01, 0.05], *trial x dialect:* β = −0.01, 95% CI = [−0.04, 0.02], *trial x surprisal x dialect:* β = −0.01, 95% CI = [−0.04, 0.02]).

3.2.2.2 Response time (Experiment 1)

The distributions of participant-specific response times for each condition are presented in Figure 3.3. Reliable main effects were observed for surprisal, dialect, and the interaction between surprisal and dialect. Response times were reliably slower for high-surprisal than low-surprisal sentences (*surprisal:* β = 0.21, 95% CI = [0.13, 0.30]); they were also slower for sentences spoken in Chengdu Mandarin than in Standard Mandarin (*dialect:* β = 0.13, 95% CI = [0.02, 0.25]). The interaction between surprisal and dialect also reliably modulated the contrast in response times between high and low surprisal conditions within each dialect: this difference was enhanced for Standard Mandarin, and slightly diminished for Chengdu Mandarin (*surprisal x dialect:* β = −0.13, 95% CI = [−0.21, −0.05]). Notably, the magnitude of the main surprisal effect exceeded its interaction with dialect, indicating listener sensitivity to the high-low surprisal contrast even in the unfamiliar Chengdu Mandarin. Based on the transformed marginal means, the estimated mean difference between high and low conditions for Chengdu Mandarin was approximately 297 ms, whereas for Standard Mandarin it was about 875 ms. While the surprisal effect was substantially larger for Standard Mandarin than Chengdu Mandarin, high surprisal nevertheless led to reliably longer response times in both dialects.

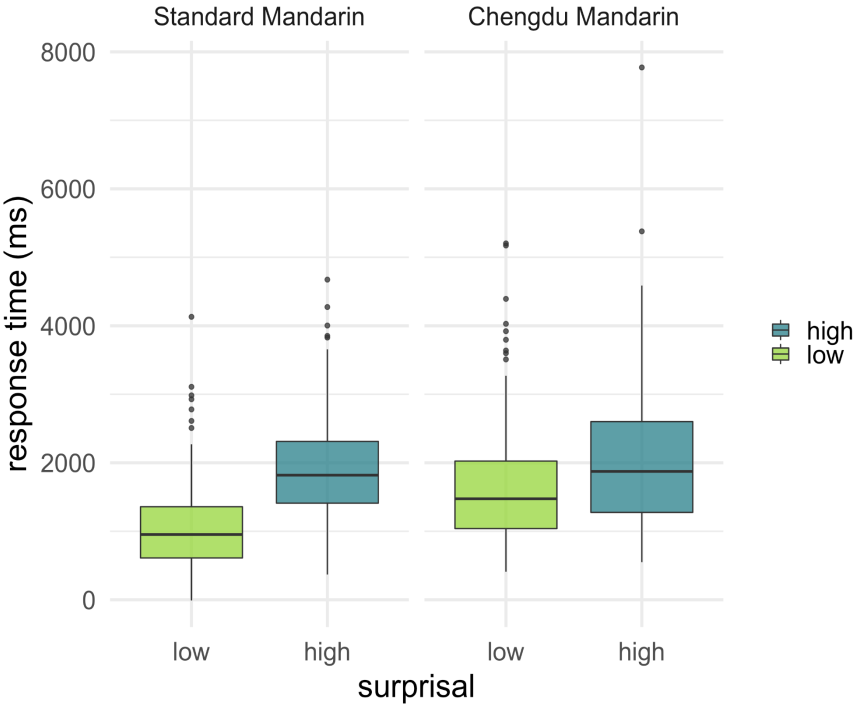


Figure 3.3 Response times across dialect and surprisal conditions.

The remaining effects of trial and its interactions with surprisal and dialect were not reliable in their direction of influence (*trial:* β = 0.0032, 95% CI = [−0.0005, 0.0067]; *trial x surprisal:* β = −0.0006, 95% CI = [−0.0025, 0.0014]; *trial x dialect:* β = 0.0007, 95% CI = [−0.0013, 0.0027]), except for the interaction between trial, surprisal and dialect (*trial x surprisal x dialect:* β = 0.0021, 95% CI = [0.0001, 0.0041]). Though response times decreased in the Standard high-surprisal condition, particularly in the initial trials, the marginal means indicate that the interaction is driven by a reliable slowdown in the Chengdu high-surprisal condition over the course of the experiment.

3.2.3 Discussion

This study investigated the relative weighting of top-down and bottom-up information in processing familiar and unfamiliar tone systems. Our findings suggested that speakers seem to attend to tone even if they do not always use it in determining word identity. When tone information was reliable, listeners correctly detected semantic implausibility, suggesting that they attend to tone even when the context introduces a strong bias against a particular lexical item. However, when tone information was unreliable, they relied on the sentential context in making their decisions. Interestingly, in both cases, response times were longer for sentences containing tones that would increase sentence surprisal, indicating that listeners are sensitive to tone even if they do not use it in determining lexical identity.

Specifically for the familiar tone system, accuracy results suggested that listeners have strong representations of segments and tones and were therefore able to use the native acoustic information, together with sentential context, to achieve high accuracy in the semantic plausibility judgment task. Moreover, response time results revealed listeners' sensitivity to the surprisal manipulation using both bottom-up and top-down information; the slowdown for the less expected lexical item indicated a strong contextual influence on lexical access.

For the unfamiliar Chengdu speech, accuracy results suggested an overriding effect of top-down information in determining sentence meaning as the listeners' judgments were overwhelmingly biased towards semantically plausible sentences based on the sentential context alone, despite any mismatch in tone. Low accuracy for the high-surprisal Chengdu sentences indicated a major bottom-up failure in identifying unexpected tones of the unfamiliar tone system. However, the overall lower accuracy for Chengdu speech compared with Standard Mandarin does not denote difficulty in understanding Chengdu Mandarin in general. Listeners consistently understood Chengdu speech well enough to correctly judge plausible sentences; they simply under-valued tone information in high-surprisal environments. For any discrepancies between observed and expected tones in the unfamiliar tone system, sentential context (i.e., top-down information) overwhelmingly guided lexical access towards a plausible judgment.

With respect to response times, a slowdown in response times in the high-surprisal condition was present to a reliable degree in both Standard Mandarin and Chengdu Mandarin. Though the magnitude of the surprisal effect was indeed greater for Standard than Chengdu Mandarin, the presence of the overall effect revealed listeners' statistically equal sensitivity to the implausibility indicated by a high-surprisal tone in both familiar and unfamiliar speech. This suggested that listeners credibly attended to the bottom-up tone information, even in unfamiliar systems, despite their ultimate bias towards a response of semantic plausibility in the Chengdu Mandarin condition.

Differences in response times across surprisal conditions indicated an unexpected integration of bottom-up information in lexical access for the unfamiliar speech. This contrasts with Gao et al. (2019)'s proposal that tone information is processed only if necessary. The Reverse Accessing Model (Gao et al., 2019) predicts that tonal information is accessed only when the discrimination is between words with different syllables in a reactivation process through "mental replay of the perceived word". In our study, the contrast was in the tone category alone with identical syllables for the target words, but listeners seemed capable of retuning the tone category–contour mapping to the extent that the measured response times were reliably different between surprisal conditions across all trials. It is likely that bottom-up processing of the novel tone acoustics happens automatically and is neither tied up with accurate lexical decision, nor induced by segmental contrast.

One plausible interpretation of this bottom-up process is that listeners extract lexical tone information from the cumulative one-minute incidental exposure of the sentences in the unfamiliar dialect. This newly received information could then be used to update the mappings between the phonological tone category and its corresponding phonetic realisation. As for the low accuracy in the semantic plausibility judgment, listeners may have been less confident about the novel tone acoustics in the unfamiliar speech, and therefore top-down information overrode the output of tone-level processing to access the sentence meaning. Although listeners failed to report the tone mismatch for the high-surprisal sentences in the unfamiliar speech, they somehow constructed tone representations using bottom-up information and responded differently in terms of response time. It is unclear whether listeners build long-lasting representations of the dialect- or talker-specific tone–contour mappings, or only temporary, task-specific representations. Whether the found adaptation persists over time and develops into perceptual learning needs to be further tested, but the awareness of the surprisal contrast, as suggested in the response time data, indicates certain degree of online adaptation.

Additionally, for the timing of online adaptation, the reliable slowdown over the course of the experiment in the high-surprisal Chengdu condition suggested gradually raised attention and awareness of bottom-up information. Critically, the lack of credibility of trial and its interactions with dialect or surprisal suggested an early-on slowdown in the high-surprisal conditions for both familiar (Standard) and unfamiliar (Chengdu) speech. This indicated that listeners may be very rapidly learning or adapting to a novel tone category–contour mapping, possibly as soon as the experiment commenced.

3.3 Experiment 2: Conditions on adaptation to an unfamiliar lexical tone system—the role of quantity and quality of incidental exposure

The follow-up experiment investigated aspects of quality and quantity of incidental exposure in adaptation to a novel lexical tone system. Chengdu Mandarin has the same underlying four-tone system as Standard Mandarin, but disparate phonetic tone realisations

Hou, 2022; Li, 2002; Zhao & Chodroff, 2022). The previous study in Section 3.2 found that native Standard Mandarin listeners adapted to a novel lexical tone system from the Chengdu Mandarin dialect with less than two minutes of incidental exposure. In that experiment, listeners consistently slowed down for the high-surprisal sentences in Chengdu Mandarin starting from the beginning of the experiment, which strongly indicated rapid adaptation to the unfamiliar phonetic tones without explicit training additional to the experimental trials.

A critical aspect in the design of the first experiment was that participants heard both the low- and high-surprisal versions of the sentence in each dialect, which provided minimal-pair sentences that contrasted in semantic plausibility and lexical tone category. The presentation of minimal-pair sentences may facilitate rapid adaptation to a novel tone system. To test the potential conditions for adapting to a novel tone system with incidental exposure, the present study investigated 1) whether adaptation can still be achieved when minimal pairs are removed, and 2) if increasing the amount of incidental exposure would facilitate adaptation. We removed the minimal-pair contrast in the dialect-specific stimuli and introduced three repetitions of all trials in the new experiment. Minimal-pair presentation may be necessary for adaptation, in which case, we would expect that listeners have comparable response times for Chengdu Mandarin between high- and low-surprisal sentences, at least in the initial trials. However, a difference in response time may still emerge as incidental exposure increases with repetition.

The following study first investigated the effect of repetition on adaptation without minimal-pair presentation via the main experiment. To single out the effect of minimal-pair presentation, the response-time data in the first repetition block from the present experiment was then extracted and compared to that from the previous experiment. The absence of minimal-pair sentences was expected to impede adaptation. However, the effect of nonminimal-pair presentation was expected to be overcome by increasing the amount of ambient exposure through repetition.

3.3.1 Methods

The current experiment replicated the design of the previous experiment of Standard Mandarin and Chengdu Mandarin, except that the sentence stimuli were selected such that no minimal pair sentences were present within either dialect condition. The trials were also repeated in three repetition blocks in the new experiment.

3.3.1.1 Participants

Thirteen native speakers of Standard Mandarin who reported little or no knowledge of Chengdu Mandarin participated in the experiment. No participant reported hearing or reading impairments.

3.3.1.2 Materials

The experiment used the same twenty-four sentence frames as in the previous experiment. For each sentence frame, the lexical tone category of one target word was manipulated to have a semantically plausible (high-surprisal) or an implausible (low-surprisal) meaning (see example sentences in Table 3.1). To avoid having both low- and high-surprisal versions of the sentence presented in the same dialect, each surprisal version was assigned to a different dialect; two lists of sentences were created differing in the dialect order. In addition, each critical-word tone category and position (medial or final, balanced evenly) were presented approximately the same number of times in each dialect.

3.3.1.3 Procedure

The experiment was built using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Participants were asked to complete the sentential semantic plausibility judgment task on their personal device with internet access in a quiet room, and with headphones if possible. They were briefed on the purpose and content of the experiment. The participants were made aware that they would be listening to sentences spoken in either the familiar Standard Mandarin or an unfamiliar Mandarin dialect. Then they were presented with a test audio so they could adjust the volume of the sound output to a comfortable level.

The practice phase consisted of two trials that introduced the test phase procedure. To start the trial, the participant pressed a “play audio” button on the screen. After listening, participants responded to the question “Does this sentence make sense” by clicking on either the “yes” or “no” response button on the screen. Feedback was then provided by displaying the orthographic form of the sentence on the screen. The two trials contained one high-surprisal and one low-surprisal Standard Mandarin sentence. These sentences were not repeated in the main experiment.

In the test phase, the participants received one of the two lists described in the materials. The task was identical to the practice trials except that no feedback was given. There were 24 trials in each block, and a total of three blocks for each dialect (24 trials × 3 repetitions × 2 dialects). The participants heard all the Chengdu Mandarin blocks first, then the Standard Mandarin blocks to avoid task-based learning effects in the adaptation process due to prior familiarity with the sentence frames in the native dialect if heard first. The trials in each repetition block were randomised.

3.3.1.4 Data analysis

The effects of *surprisal* (high surprisal vs. low surprisal), *dialect* (Standard Mandarin vs. Chengdu Mandarin) and *repetition* were assessed on accuracy and response time. Responses that matched the expected plausibility judgment were considered *correct*: “yes” responses to low-surprisal plausible sentences and “no” responses to high-surprisal implausible sentences. Response time was calculated as the interval between the end of the audio file and the click registering a judgment.

Accuracy was modelled with a Bayesian logistic mixed-effects regression, and response time with a Bayesian log-normal mixed-effects regression, both with weakly informative priors (Bürkner, 2018). Each model included fixed effects of surprisal, dialect, two repetition contrasts, the full set of interactions. The random effect structure for participants included an intercept and slopes for surprisal, dialect, repetition contrasts and all the interactions, and for sentence frame, an intercept, and a random slope for surprisal. The priors for the accuracy model were $N(0, 20)$ for the intercept, main effects, and interactions, and $N(0, 0.05)$ for random effects. For the response-time model, the priors were $N(7, 1)$ for the intercept, $N(0, 1)$ for main effects and interactions, and $N(0, 0.01)$ for random effects.

Since tone categories were counterbalanced across all the conditions in this experiment, the accuracy and response time models were run for a second time with the tone factors added. The fixed factors also included the interactions between dialect and each tone contrast, and the interactions of dialect, surprisal and each tone contrast; the priors were set as in the above models. The results of the tone model can be found in the Appendix g. The results reported in text for the *dialect, surprisal* and *repetition* factors were based on the initial modelling without the tone factors.

The models were run for 2000 iterations with a burn-in period of 1000 iterations. Surprisal and dialect were sum-coded (*surprisal:* high-surprisal = 1, low-surprisal = −1; *dialect:* Chengdu Mandarin = 1, Standard Mandarin = −1). Reverse Helmert coding was used for the three-level repetition factor, comparing one level to the mean of the previous level(s) (*repetition contrast 1*: block 2 = 1/2, block 1= -1/2, block 3 = 0; *repetition contrast 2*: block 3 = 2/3, block 2 = − 1/3, block 1 = −1/3). Adding in the tone factors did not affect the interpretation of the other fixed effects. For the tone factor, deviation coding was applied to compare the level of Tone 1, Tone 2, and Tone 3 to the overall mean respectively; Tone 4 was left not compared given the default coding scheme. An estimate was deemed credible in its direction of influence on the dependent variable if the 95% credible interval excluded 0 (i.e., no effect).

Additionally for the comparison between the present and previous experiments, response time was modelled with fixed effects of surprisal, dialect, trial, presentation (*presentation*: with-minimal-pair design = 1, no-minimal-pair design = −1), and the interaction of surprisal, dialect and presentation. The random effects for participants included an intercept and slopes for surprisal and dialect, and for frame an intercept and a slope for surprisal. The same priors were used as the above response-time model.

3.3.2 Results

3.3.2.1 Accuracy (Experiment 2)

The overall accuracy across dialect and surprisal conditions (Figure 3.4) closely resembled the previous findings: 1) high near-ceiling accuracy in both surprisal conditions for the familiar dialect and 2) considerably lower accuracy in the high-surprisal condition for the unfamiliar dialect, Chengdu Mandarin. A gradual improvement can be seen over the three blocks in the Chengdu high-surprisal conditions.

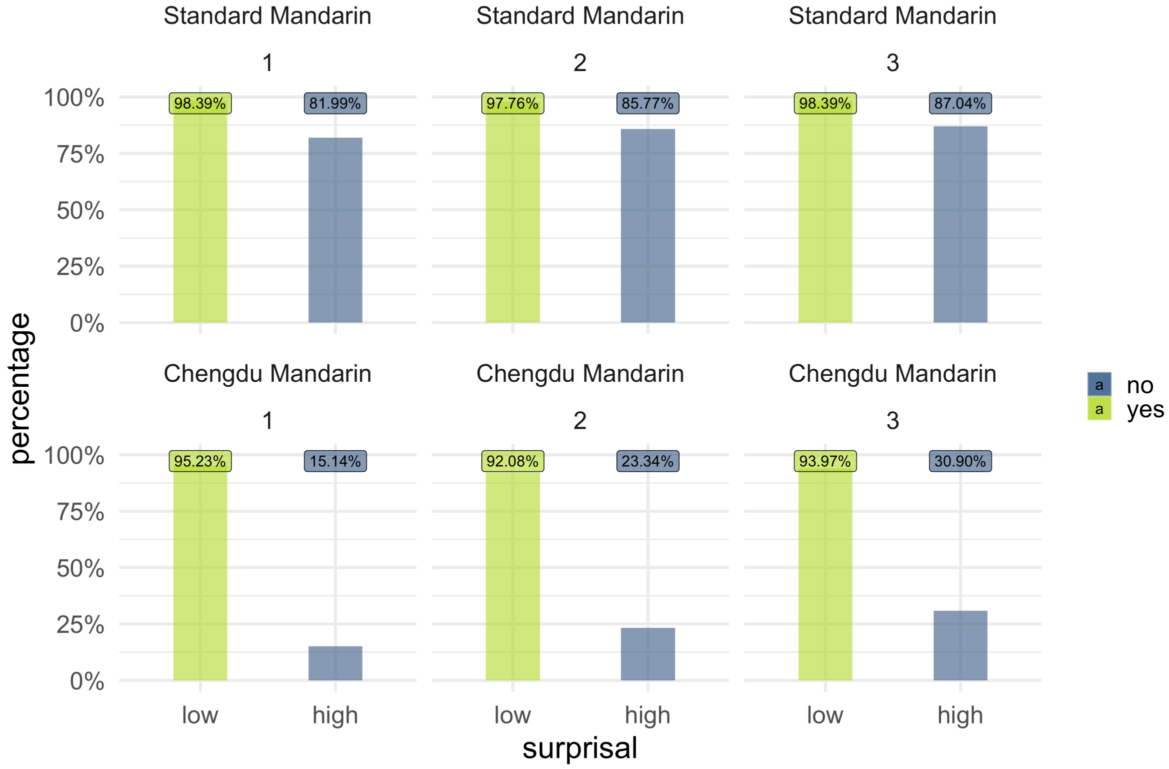


Figure 3.4 Percentage of correct responses across dialect, surprisal and repetition (“1, 2, 3” refer to the repetition blocks).

The model revealed credible main effects of surprisal, dialect, and the interaction between dialect and surprisal as in the previous study (see results in Appendix f.). Specifically, accuracy was higher in the low-surprisal condition than in the high-surprisal condition (*surprisal:* β = −7.64, 95% CI = [−14.89, −3.10]); accuracy was higher for sentences spoken in Standard Mandarin than in Chengdu Mandarin (*dialect:* β = −7.04, 95% CI = [−14.24, −2.51]); the credible interaction between surprisal and dialect indicated higher accuracy for Standard Mandarin low-surprisal sentences (*dialect x surprisal:* β = 5.13, 95% CI = [0.58, 12.32]).

For the effect of repetition, accuracy was reliably different from the first to the second block of repetition (*repetition contrast 1*: β = −7.78, 95% CI = [−21.09, −0.19]), but not from the first two blocks to the third block (*repetition contrast 2*: β = 4.79, 95% CI = [−6.06, 19.63]). The second repetition block reliably interacted with surprisal (*surprisal x repetition contrast 1*: β = 8.29, 95% CI = [0.73, 21.59]) and dialect (*dialect x repetition contrast 1*: β = 7.70, 95% CI = [0.11, 20.95]), while the third repetition block showed no credible interaction with the other factors (*surprisal x repetition contrast 2*: β = −4.23, 95% CI = [−18.96, 6.62]; *dialect x repetition contrast 2*: β = −4.39, 95% CI = [−19.14, 6.43]; *surprisal x dialect x repetition contrast 2*: β = 4.63, 95% CI = [−6.13, 19.48]). This suggested that accuracy improved after the second repetition of the trials for the high-surprisal sentences compared to the low-surprisal, and for Chengdu sentences compared to Standard Mandarin, but these did not reliably improve in the third block.

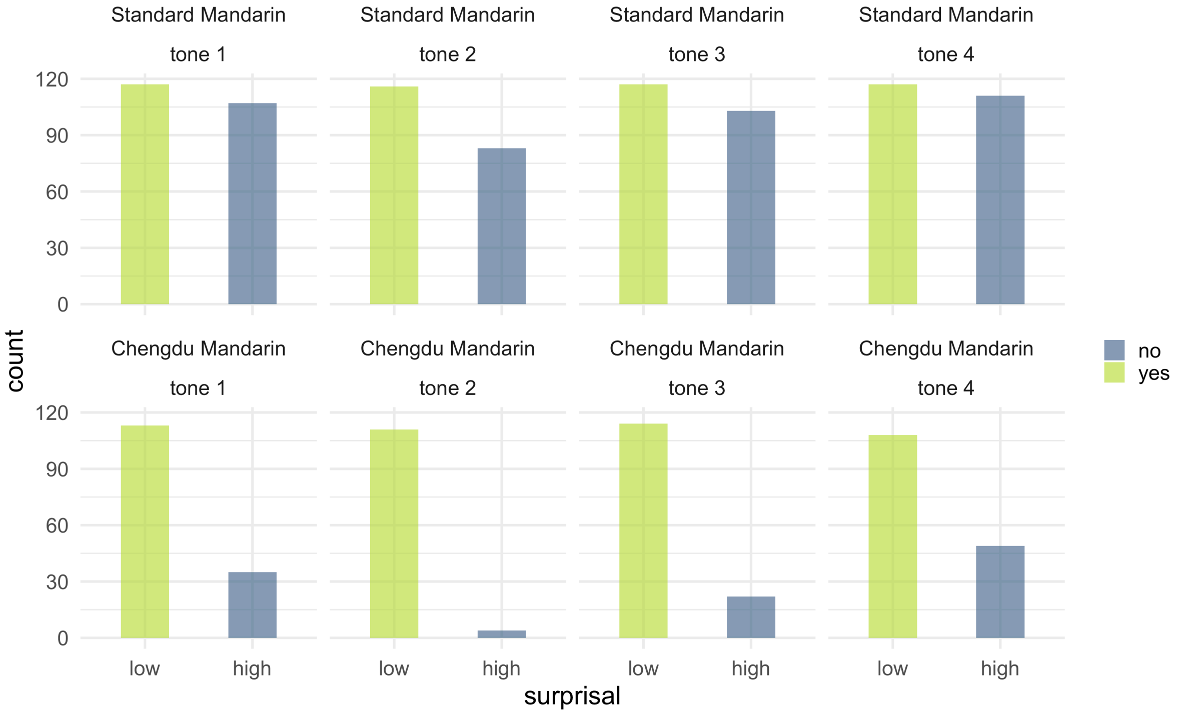


Figure 3.5 Count of correct responses across dialect, surprisal and tone conditions.

For the effect of tone category (Figure 3.5), the follow-up model did not detect credible main effect for any of the tone factors and the interactions, which indicated statistically comparable accuracy across tone categories in both familiar and unfamiliar dialect. However, numerically speaking, accuracy was particularly low for the sentences containing Chengdu Tone 2 and relatively higher for Chengdu Tone 4.

3.3.2.2 Response time (Experiment 2)

The response-time model identified credible effects of all tested factors and their interactions, except for the interaction between surprisal and the second repetition contrast (see results in Appendix f.). To be exact, the credible effects of surprisal (*surprisal:* β = 0.15, 95% CI = [0.12, 0.18]), dialect (*dialect:* β = 0.11, 95% CI = [0.08, 0.13]) and the interaction between surprisal and dialect (*surprisal x dialect:* β = −0.07, 95% CI = [−0.09, −0.05]) replicated the patterns found in the previous study: listeners were reliably slower in the high-surprisal condition in both dialects, with a greater difference between the surprisal conditions in Standard Mandarin than in Chengdu Mandarin (Figure 3.6). Nevertheless, a difference was still observed between high- and low-surprisal conditions in Chengdu Mandarin.

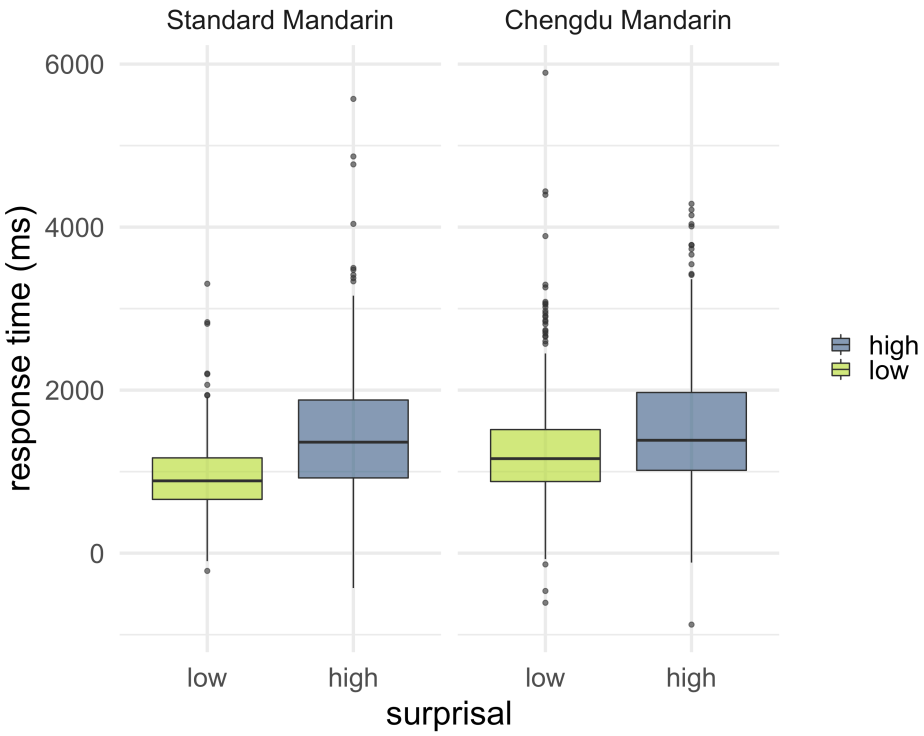


Figure 3.6 Response times across dialect and surprisal conditions.

For the effect of repetition (Figure 3.7), all responses generally accelerated block by block (*repetition contrast 1*: β = −0.16, 95% CI = [−0.21, −0.11]; *repetition contrast 2*: β = −0.19, 95% CI = [−0.24, −0.14]). Moreover, slower responses were found for Chengdu sentences after each repetition, relative to Standard Mandarin sentences (*dialect x repetition contrast 1*: β = 0.11, 95% CI = [0.06, 0.17]; *dialect x repetition contrast 2*: β = 0.10, 95% CI = [0.05, 0.14]). However, response times were credibly faster for high-surprisal sentences from the first to the second block (*surprisal x repetition contrast 1*: β = −0.06, 95% CI = [−0.11, −0.003]), possibly driven by the faster responses to Standard Mandarin sentences, but there was no difference towards the third block (*surprisal x repetition contrast 2*: β = −0.02, 95% CI = [−0.07, 0.02]). In fact, response times were reliably modulated by the three-way interactions between surprisal, dialect and both repetition contrasts (*surprisal x dialect x repetition contrast 1*: β = 0.07, 95% CI = [0.02, 0.13]; *surprisal x dialect x repetition contrast*

2: β = 0.06, 95% CI = [0.01, 0.11]), suggesting block-wise slowdown for Chengdu high-surprisal sentences, but block-wise speed-up for Standard Mandarin high-surprisal sentences.

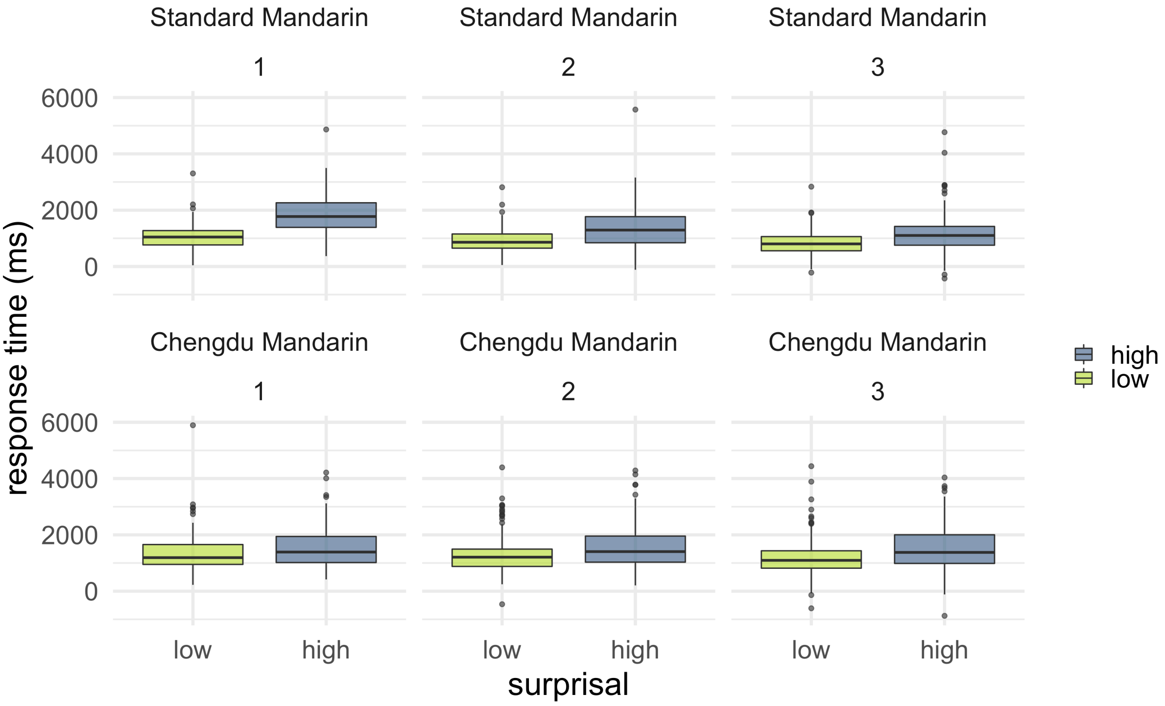


Figure 3.7 Response times across dialect, surprisal and repetition conditions ("1, 2, 3" refer to the repetition blocks).

For the effect of tone category (Figure 3.8), credible interactions were detected between dialect and Tone 1 (*dialect x Tone 1*: β = −0.09, 95% CI = [−0.13, −0.05]), suggesting faster responses to the Chengdu sentences with Tone 1 for the target word. The follow-up model also identified credible interactions between surprisal, dialect and Tone 2 (*surprisal x dialect x Tone 2*: β = −0.04, 95% CI = [−0.08, −0.004]), and between surprisal, dialect and Tone 3 (*surprisal x dialect x Tone 3*: β = 0.04, 95% CI = [0.01, 0.08]). Specifically, response times were reliably slower for the Chengdu sentences with Tone 2 in the low-surprisal condition; also, response times were reliably slower for the Chengdu sentences with Tone 3 in the high-surprisal condition. This may suggest weaker discrimination between the surprisal conditions for

Chengdu Tone 2 and greater discrimination between the surprisal conditions for Chengdu Tone 3.

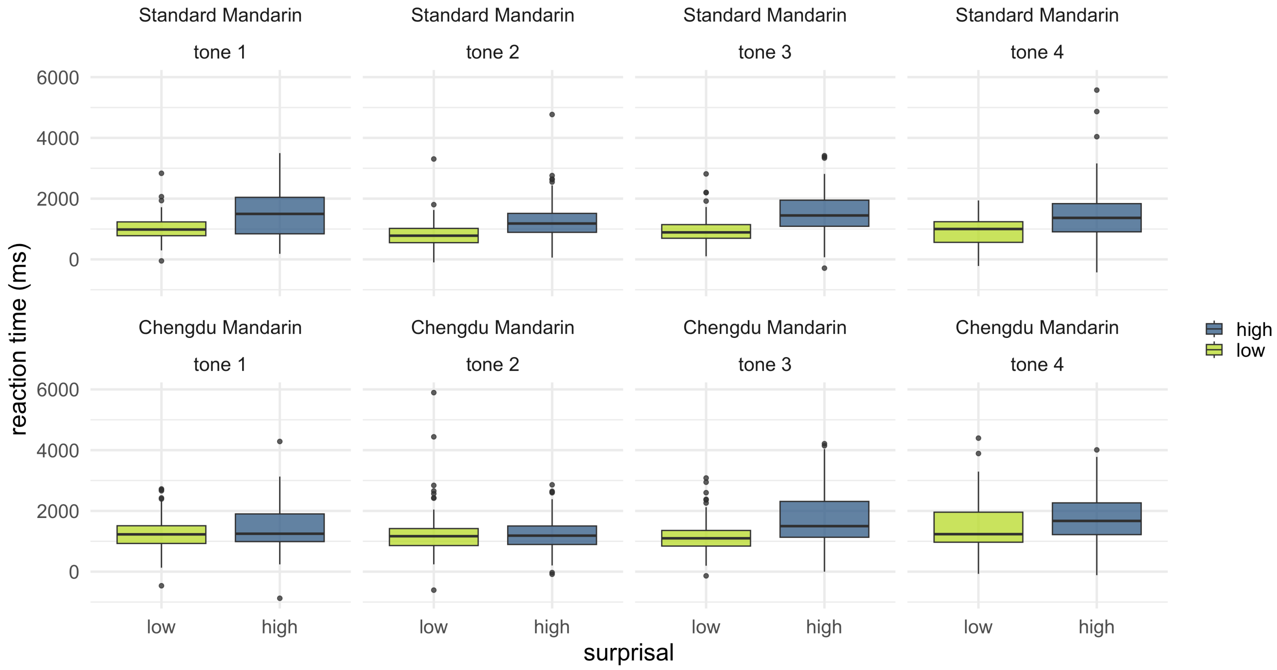


Figure 3.8 Response times across dialect, surprisal and tone conditions.

3.3.2.3 The effect of minimal pairs

To examine the effect of minimal-pair presentation on response time, we compared the first block of data in the present study to the data in the previous study (Section 3.2), which only differed in the presence of minimal pairs (Figure 3.9). Accuracy was not examined as surprisal did not influence responses to Chengdu sentences in either study.

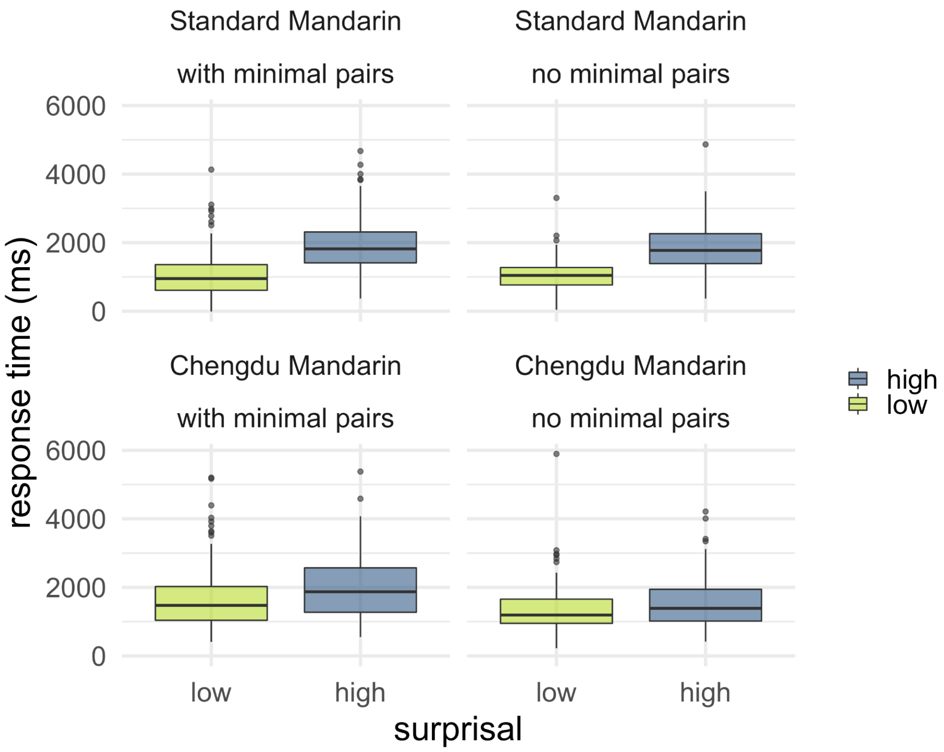


Figure 3.9 Response times across dialect, surprisal and presentation conditions in the previous (with-minimal-pair) and the new (no-minimal-pair) experiments.

According to the model, no credible effect of presentation was detected between the two designs (*presentation*: β = 0.05, 95% CI = [−0.03, 0.12]), indicating that the effect of minimal-pair presentation was not as salient as expected. Response times were consistently slower for high-surprisal sentences in both experiments (*surprisal*: β = 0.20, 95% CI = [0.15, 0.25]), reflecting listeners' awareness of surprisal manipulation even without minimal pairs or repetition. The high-surprisal slowdown did not reliably interact with presentation (*surprisal x presentation*: β = 0.01, 95% CI = [−0.03, 0.06]; *surprisal x dialect x presentation*: β = −0.0006, 95% CI = [−0.04, 0.03]), indicating that the removal of minimal pairs did not credibly affect listeners' sensitivity to the surprisal manipulation. Nevertheless, the estimated mean difference between the Chengdu high- and low-surprisal conditions differed numerically between the two sets of data in the expected direction: about 170 ms without minimal pairs, and 270 ms with minimal pairs. Minimal-pair presentation may have numerically facilitated adaptation to the novel tone system, resulting in greater distinction between the surprisal conditions; removal of the minimal pairs reduced, but did not obviate the effect of surprisal. There was also a credible interaction between presentation and dialect (*dialect x presentation*: β = 0.06, 95% CI = [0.003, 0.11]), indicating slower responses to Chengdu sentences when minimal pairs were present.

3.3.3 Discussion

The previous experiment on Chengdu Mandarin revealed rapid adaptation to the novel tone system with incidental exposure containing minimal-pair sentences (Section 3.2). The current experiment found that rapid adaptation to the novel tone system was persistent even when minimal-pair sentences were removed from the stimuli and only minimal incidental exposure was available. Enhancement in quality (minimal pairs) and quantity (repetition) of the exposure both facilitated adaptation in terms of accuracy and response time, but they were not necessary factors.

The present study showed similar results for the effects of dialect and surprisal as in the previous study, which indicated successful adaptation and shared perceptual mechanisms for novel tone processing with and without minimal pairs. Increased incidental exposure reliably boosted adaptation to the unfamiliar tone system, evidenced by improved accuracy and increased difference in response times between the low- and high-surprisal conditions showing over the second block of repetition.

Specifically for the effect of minimal-pair presentation, it was surprising to find that listeners were sensitive to the surprisal manipulations even when the minimal pairs were removed from exposure. In fact, adaptation occurred under rather adverse conditions, where incidental exposure was limited to one repetition and with no minimal-pair sentences in the same dialect. Nevertheless, inclusion of minimal pairs in the stimuli can assist discrimination between the low- and high-surprisal meanings, and may potentially direct more attention to the tone contrast and ease the process of adaptation or learning of the new tone system.

As minimal incidental exposure was sufficient for adapting to the unfamiliar tone system with or without minimal pairs in the exposure, it is unlikely that listeners relied on increased exposure over one minute to initiate adaptation. Repetition in this experiment was more likely a consolidating factor as the mappings between the phonological categories and the novel phonetic tone realisations were reinforced with more available information.

3.4 Conclusion

To conclude, the study on top-down and bottom-up processing of tone systems tested the relative role of tonal information for sentence interpretation in two Mandarin dialect varieties. We found that contrary to the previous expectation, phonetic tonal information seems to be processed, even when its mapping to the phonological tone categories is unfamiliar. This finding also leads to the hypothesis that phonetic tone information is always processed — even if such information may have little influence in lexical decision (Gao et al., 2019). This has broader implications for models of speech perception involving lexical tone (Strauss, Harris & Magnuson, 2007; Shuai & Malins, 2017; Gao et al., 2019). Further research would be needed to address dialect- and tone-specific perception with carefully balanced tone contrast for a broader range of Mandarin group dialects other than Chengdu Mandarin. The current experimental design with the surprisal and dialect manipulations could also be extended by introducing an exposure phase to the unfamiliar dialect to explore perceptual adaptation to or learning of unfamiliar tone systems with explicit training.

The follow-up study investigated the potential factors impacting adaptation to an unfamiliar tone system by testing the effects of minimal-pair presentation of the stimuli and increased incidental exposure. We found that both factors were adequate to induce adaptation, though neither was necessary. It also has implications of more readily processed tones for models of speech perception. Further analysis should delve into the tone-specific adaptation. The experiments in this chapter examined adaptation with incidental exposure directly from the experimental trials; the following Chapter 4 assessed the effect of explicit training for adaptation to unfamiliar lexical tone systems.

3.5 Authorship and publication status

This chapter combines and reorganises the content published in the two peer-reviewed papers: "Top-Down and Bottom-up Processing of Familiar and Unfamiliar Mandarin Dialect Tone Systems" in the Proceedings of Speech Prosody 2022 (Zhao, Sloggett & Chodroff, 2022) and "Conditions on Adaptation to an Unfamiliar Lexical Tone System: The Role of Quantity and Quality of Exposure" in the Proceedings of the 20th International Congress of Phonetic Science (Zhao, Sloggett & Chodroff, 2023). Both papers were co-authored by Liang Zhao, Shayne Sloggett, and Eleanor Chodroff. The relative contribution is as follows:

Liang Zhao: Conceptualisation, methodology, investigation, resources, data curation, formal analysis, writing—original draft, writing—review & editing, visualisation; Shayne Sloggett: methodology, formal analysis, writing—review & editing; Eleanor Chodroff: methodology, visualisation, formal analysis, writing—review & editing, supervision, funding acquisition.

Chapter 4 Perception of lexical tone variation with explicit exposure

4.1 Introduction

According to the findings in Chapter 3, listeners readily accommodate dialectal variation in lexical tone systems through incidental exposure. Chapter 4 examined whether adaptation can be further facilitated with explicitly introduced passage exposure of the unfamiliar dialect, and how the similarity of the dialectal tone system to the native tone system would modulate adaptation performance. Chengdu Mandarin and Jinan Mandarin both have a four-tone system, but differ in their relative similarity of the phonetic tone inventory to the Standard Mandarin tone system: Jinan Mandarin tones are phonetically more similar to Standard Mandarin than Chengdu Mandarin tones. We experimented on native Standard Mandarin speakers' adaptation to Chengdu Mandarin and Jinan Mandarin tones before and after explicit passage exposure. A similar version of the sentence semantic-plausibility judgment task used in Chapter 3 was implemented in two consecutive experiments which contrasted in the presence (Experiment 3) or absence (Experiment 4) of minimal-pair sentences in the stimuli.

The following introduction sections briefly introduce adaptation with explicit training and clarifies the phonetic similarity of Chengdu and Jinan tone systems relative to Standard Mandarin. Sections 4.2 and 4.3 illustrate the methods and results of Experiment 3 and Experiment 4 respectively. Since previous findings negated the necessity of minimal-pair sentence presentation in initiating adaptation to the unfamiliar tone system (Chapter 3), in Section 4.4 we combined the data from the two experiments in order to assess the overall effect of explicit exposure and dialectal tone similarity on adaptation to unfamiliar Mandarin dialect tone systems. Section 4.5 provides the conclusion of the major findings and implications to current understanding of lexical tone processing.

4.1.1 Adaptation with increased exposure

Adaptation to unfamiliar speech can be conditioned by various factors (see more detail in Section 1.4.3). Though former findings suggest that adaptation can be reasonably successful with incidental exposure in a short period (Clarke & Garrett, 2004; Chapter 3), increased exposure may help listeners to better generalise heard patterns to novel sounds or novel speakers.

Either the increased amount exposure from one specific speaker or the inclusion of more speakers in the stimuli would presumably contribute to more sufficient input for potential adaptation. In Maye et al.'s study (2008), successful adaptation to a modified English accent from a specific talker was found after a relatively long period of passage exposure about twenty minutes. For accented-speech, significant improvement was detected in the third quartile of the sentence stimuli out of four chunks in total (Bradlow & Bent 2008). However, when speech of multiple speakers is presented, the results are somewhat mixed. Bradlow & Bent's study (2008) showed that exposure to multi-speaker utterances greatly facilitated talker-independent adaptation. Nevertheless, Floccia et al. (2006) argued that exposure to multiple speakers could make adaptation even more difficult since listener's attention was directed to cross-talker differences, rather than similarities. It is also assumed that if presented with single-talker stimuli, longer exposure would be expected to initiate significant improvement in perceiving non-native speech (Bradlow & Bent 2008).

Research on the exposure-induced adaptation has focused on variation on the segmental level; not much research has been done on the level of lexical tones. What happens when the unfamiliar accent mostly targets the phonetic realisations of lexical tones? This is exactly the case with Mandarin dialects. With explicit passage exposure to the unfamiliar Mandarin dialects, we expected improved adaptation for both Chengdu and Jinan Mandarin with higher accuracy and faster responses in the sentence plausibility judgment task.

4.1.2 Phonetically similar vs. phonologically similar tone system

Apart from the amount of exposure, the degree of variability of the heard speech relative to the listeners' native speech may also affect how well and how quickly the unfamiliar phonetic realisations can be processed and adapted to. The Perceptual Assimilation Model (Best, 1994) suggests that greater dissimilarity to the native speech leads to easier discrimination or better perception (So & Best, 2010; Reid et al., 2015). Relative to the Standard Mandarin tone system, Jinan Mandarin has comparable acoustic and especially perceptual phonetic contours, but highly disparate tone–contour mappings (Figure 4.1). The Jinan tone inventory contains all the contour types, i.e. level, rising, dipping and falling, that are present in the Standard Mandarin tone system. Specifically, Jinan Mandarin Tone 2 phonetically resembles Standard Mandarin Tone 1 (the level tone); and Jinan Tone 4 phonetically resembles Standard Mandarin Tone 4 (the falling tone). In contrast, Chengdu Mandarin has fairly disparate phonetic contours compared to Standard Mandarin, and critically no level tone. Given the greater dissimilarity between the Chengdu and Standard Mandarin tone systems than between the Jinan and Standard Mandarin, we expected listeners' overall better perception of Chengdu Mandarin speech than Jinan Mandarin speech.

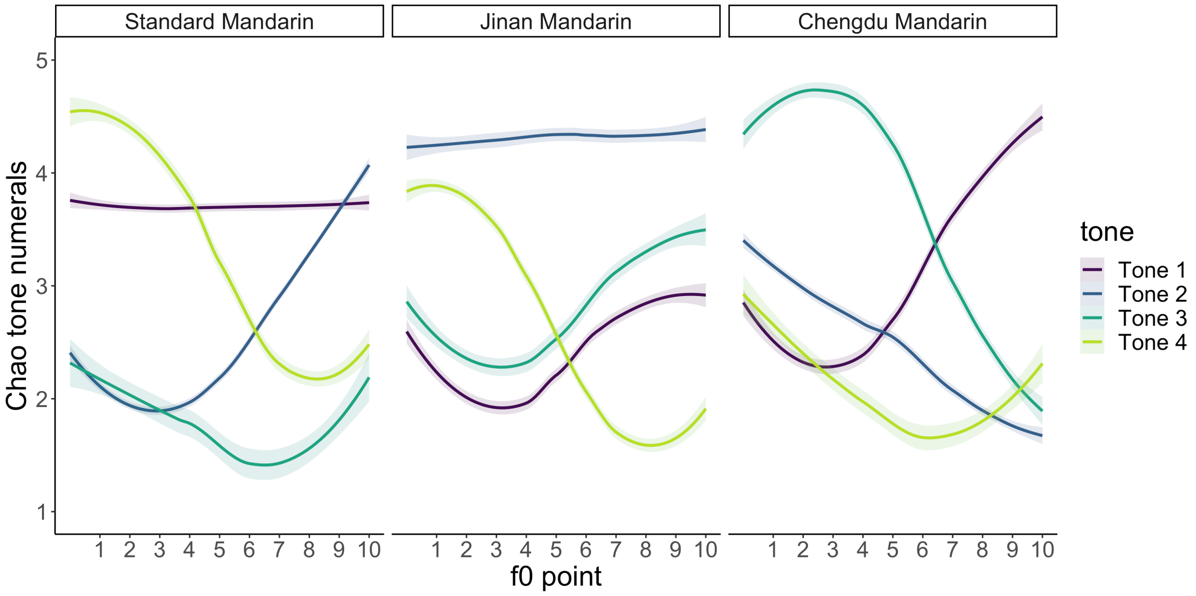


Figure 4.1 Smoothed lexical tone contours of Standard, Jinan and Chengdu Mandarin converted to Chao Tone numerals (Zhao & Chodroff, 2022). Ribbons reflect ± 0.2 standard error of the mean.

4.2 Experiment 3: Adaptation to Chengdu vs. Jinan tones with minimal pairs

Experiment 3 used the same set of surprisal sentences as in Chapter 3, by which the reliability of tonal information of a target word was manipulated in high and low surprisal conditions. A 2×2×2 factorial design was used to assess the effects of sentence semantic plausibility (high surprisal vs. low surprisal), passage exposure (pre-exposure vs. post-exposure) and dialect (Chengdu Mandarin vs. Jinan Mandarin) on the accuracy of semantic plausibility judgments and response times. Both high-surprisal and low-surprisal versions of the sentence were presented in each condition and listeners heard minimal-pair sentences in the assigned dialect.

4.2.1 Methods

4.2.1.1 Participants

Twenty native speakers of Standard Mandarin who reported little or no knowledge of Chengdu or Jinan Mandarin participated in the experiment. No participant reported hearing or reading impairments. Participants were randomly assigned to either the Chengdu task or the Jinan task.

4.2.1.2 Materials

We used the same twenty-four sentence pairs manipulated in high or low surprisal as in Chapter 3 in a between-item design. Participants heard twelve pairs of sentence item in the pre-exposure phase and different twelve sentence pairs in the post-exposure phase. Half the critical words were sentence-medial and half were sentence-final. An example sentence pair can be found in Table 3.1 in the previous chapter. Participants heard both renditions of each sentence for a total of 48 trials (24 items × 2 surprisal conditions). For each dialect, 48 sentence recordings were prepared; participant heard either the Chengdu set of stimuli, or the Jinan set of stimuli.

The exposure passage was *The North Wind and the Sun* translated in Simplified Standard Mandarin. The exposure recordings were recorded by a male native speaker of Chengdu Mandarin (aged 29) and a female native speaker of Jinan Mandarin (aged 27). The speakers were instructed to read aloud the written translation of *The North Wind and Sun* clearly and fluently for the recordings. The passage recordings were also augmented into separate sentences as part of stimuli in the exposure phase. The surprisal sentence recordings were also produced by the same two speakers. A 10-ms silence was inserted at the beginning of each sentence, and the audio file was scaled to an intensity of 70dB.

4.2.1.3 Procedure

The experiment was hosted using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Participants were asked to complete the sentential semantic plausibility judgment task on a device with internet access in a quiet environment, and with headphones if possible. They were first briefed on the purpose and content of the experiment. The participants were made aware that they would be listening to sentences spoken in some unfamiliar Mandarin dialects. Then they were presented with a test audio and adjusted the volume of the sound output to a comfortable level.

In the practice phase, participants listened to two example pairs of high-surprisal and low-surprisal sentences in Standard Mandarin, answered the question "*Does this sentence make sense*", and then received feedback regarding their answer in the form of a written version of the sentence. Specifically, participants were instructed to click the "play" button to start the audio and then click on the "yes" or "no" button on the screen to answer the question. The correct answer and the sentence in simplified Chinese characters were then presented. The "yes" and "no" buttons were presented closely adjacent to each other at the centre of the screen with the "yes" button on the left side and the "no" button on the right side.

In the test phase, the participant received either the Chengdu task or the Jinan task. The presentation of trials in each exposure condition was fully randomised. The required response was identical to the familiarisation stage except that no feedback was provided. Participants first heard sentences in the pre-exposure phase. Then, they were instructed to listen to the passage exposure twice. In the first time of listening to the exposure passage, participants were asked to listen to the whole story carefully with no written form on the screen. In the second time of listening, the passage was played sentence by sentence; each sentence was presented in Simplified Standard Mandarin on the screen with two key words replaced by blanks. The selected key words were noun words with meanings closely related to *the North Wind and the Sun* story (e.g., "*tai4 yang2*" for the word "sun"; "*pao2 zi5*" for the word "coat"). Participant were asked to fill in the blanks after hearing each individual sentence. Following the exposure phase, participants underwent the same task of sentence semantic plausibility judgment in the post-exposure trials.

4.2.1.4 Data analysis

Accuracy and response time from the test phase were analysed as dependent variables across manipulations of surprisal (high-surprisal vs. low-surprisal), exposure (pre-exposure vs. post-exposure) and dialect (Chengdu Mandarin vs. Jinan Mandarin).

"Yes" responses to low-surprisal (i.e., plausible) sentences and "No" responses to high-surprisal (i.e., implausible) sentences were coded as "correct" responses. Response time was calculated as the interval between the end of the audio file and the click registering a judgment. Ten trials were excluded from the analysis due to missing or negative response times. This left a total of 470 trials for analysis.

Accuracy was modelled with a Bayesian logistic mixed-effects regression, and response time with a Bayesian log-normal mixed-effects regression, both with weakly informative priors (Bürkner, 2018). Each model included fixed effects of surprisal, exposure, dialect, trial number, and the full set of interactions. The random effect structure for participant included an intercept and slopes for surprisal, dialect, trial number and the interaction between surprisal and dialect, and for sentence frame, an intercept and random slope for dialect. Priors for main effects and interactions were normal distributions centred on 0 with a standard deviation of 20 for the accuracy model and log-normal distributions centred on 0 with a standard deviation of 1 for the response time model. The prior for the intercept was $\mathcal{N}(0, 20)$ for accuracy and $\mathcal{N}(7, 1)$ for response time; for the random effects, it was $\mathcal{N}(0, 0.05)$ for accuracy and $\mathcal{N}(0, 0.01)$ for response time. The model was run for 2000 iterations with a burn-in period of 1000 iterations. Surprisal, exposure and dialect were sum-coded (surprisal: high-surprisal = 1, low-surprisal = -1; exposure: post-exposure = 1, pre-exposure = -1; dialect: Chengdu Mandarin = 1, Jinan Mandarin = -1), and trial number was centred on the mean. If the 95% credible interval for an estimated effect excluded 0 (i.e., no effect), then it was deemed to be credible in its direction of influence on the respective dependent variable.

4.2.2 Results

4.2.2.1 Accuracy (Experiment 3)

As shown in Figure 4.2, accuracy was around 90% for low-surprisal conditions in both Chengdu and Jinan Mandarin in the pre- and post-exposure phase, but dropped considerably in the high surprisal conditions for both dialects. The participants' judgment was biased towards a plausible "yes" response for both low- and high-surprisal sentences in the two dialects, which led to higher accuracy for low-surprisal plausible sentences and lower accuracy for high-surprisal implausible sentences as expected.

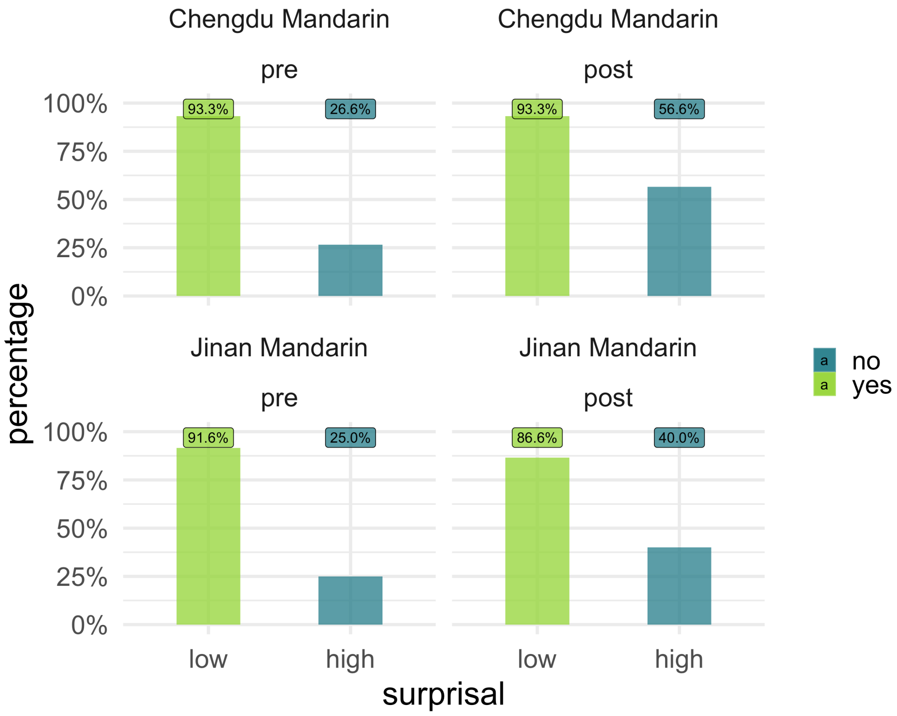


Figure 4.2 Percentage of "correct" responses across dialect, exposure and surprisal conditions. "Yes" (plausible; lime green) is treated as the correct response for low-surprisal conditions, and "no" (not plausible; blue green) as correct for high-surprisal conditions.

According to the model, there were credible main effects of surprisal, dialect, the interaction between dialect and exposure, and the interaction between exposure and trial on accuracy. Specifically, accuracy was higher in the low-surprisal condition than in the high-surprisal condition (surprisal: β = −2 .96, 95% CI = [−4.18, −1.91]). In addition, accuracy was higher for the sentences spoken in Chengdu Mandarin than in Jinan Mandarin (dialect: β = 1.04, 95% CI = [0.26, 1.96]). A credible interaction was observed between dialect and exposure, indicating higher accuracy after exposure for the Chengdu sentences than the Jinan sentences (dialect x exposure: β = 0.79, 95% CI = [0.13, 1.62]). A credible interaction was also found between exposure and trial (exposure x trial: β = 0.07, 95% CI = [0.01, 0.13], suggesting greater improvement over trials after the exposure than before the exposure.

Although the model did not reveal credible main effect of exposure (exposure: β = 0.98, 95% CI = [−0.08, 2.15]), and trial (trial: β = 0.06, 95% CI = [−0.01, 0.13]). Figure 4.2 illustrates the numerically increased accuracy after exposure, particularly for the high surprisal sentences in both dialects; and such improvement seemed greater in the Chengdu condition than the Jinan condition. To examine the effect of exposure in Chengdu and Jinan Mandarin respectively, we removed the fixed factor of dialect and ran separate models on the Chengdu data and the Jinan data. The results showed a credible main effect of exposure in the Chengdu model (exposure in Chengdu: β = 2.25, 95% CI = 0.43, 4.59], but no credible effect of exposure in the Jinan model (exposure in Jinan: β = 0.47, 95% CI = [−1.21, 2.32]). This further indicated that the explicit exposure reliably increased accuracy for the Chengdu sentences, but not for the Jinan sentences, although accuracy for the Jinan sentences was numerically higher after the exposure, especially in the high-surprisal condition (Figure 4.2).

4.2.2.2 Response times (Experiment 3)

Figure 4.3 demonstrates the response times across dialect, exposure and surprisal conditions. Reliable main effects were observed for surprisal, the interaction of surprisal, dialect and exposure, and the interaction of dialect, exposure and trial. Response times were reliably slower for high-surprisal than low-surprisal sentences across dialect and exposure conditions (surprisal: β = 156.62, 95% CI = [24.59, 289.52]), suggesting participants' awareness of high-surprisal manipulation for the sentences in both dialects. However, no credible effect was found for dialect (dialect: β = −93.67, 95% CI = [−337.52, 143.31]), exposure (exposure: β = −24.43, 95% CI = [−150.87, 102.14]) and trial (trial: β = 6.40, 95% CI = [−4.02, 16.85]), which indicated listeners' overall equal sensitivity to the high-surprisal sentences with or without exposure, in both dialects, and throughout the experiment.

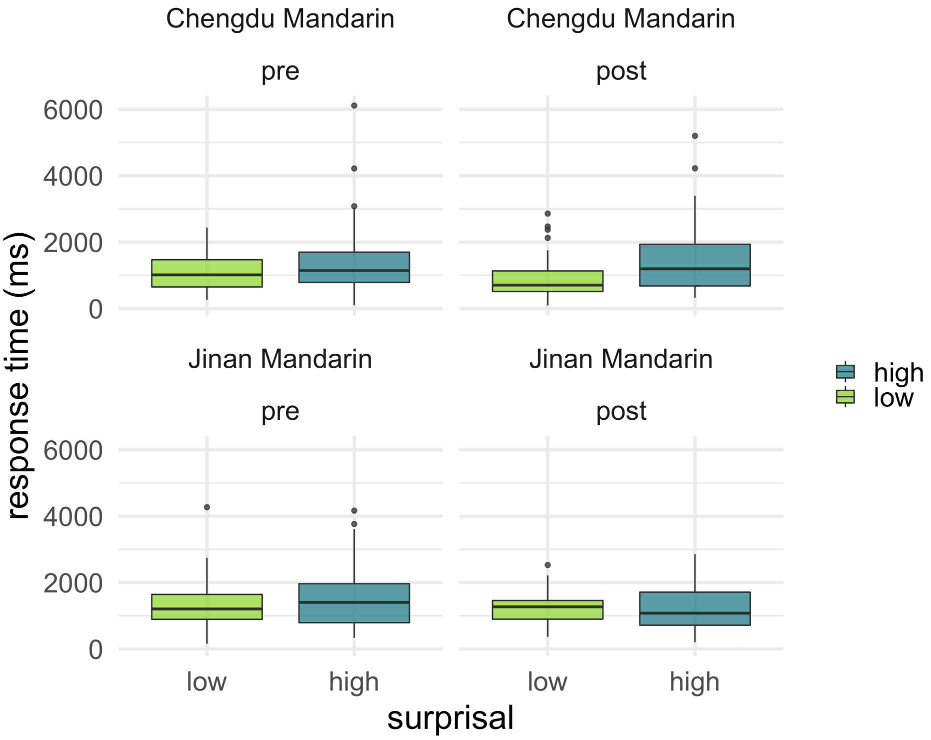


Figure 4.3 Response times across dialect, exposure and surprisal conditions in Experiment 3.

The interaction of surprisal, dialect and exposure reliably modulated the contrast in response times between pre- and post-exposure conditions for high-surprisal sentences in each dialect (surprisal x dialect x exposure: β = 92.01, 95% CI = 23.65, 163.07]). The response time was reliably longer after exposure in the Chengdu high-surprisal conditions, while it was shorter after exposure in the Jinan high-surprisal conditions. The estimated mean difference between the pre- and post-exposure Chengdu high-surprisal conditions was about 224.36 ms, and that between the pre- and post-exposure Jinan high-surprisal conditions was −218.68 ms.

A credible interaction was also found between dialect, exposure and trial (dialect x exposure x trial: β = −8.24, 95% CI = [−15.30, −1.13]), which suggested faster responses to the Chengdu sentences over trials after the exposure than to the Jinan sentences. The other interactions of exposure with either surprisal, dialect or trial were not statistically reliable in their directions of influence (exposure x dialect: β = 18.75, 95% CI = [−53.42, 92.33]; exposure x surprisal: β = 25.28, 95% CI = [−67.26, 123.76]; exposure x trial: β = 6.03, 95% CI = [−1.13, 13.21]).

4.2.3 Discussion

Experiment 3 investigated adaptation to Chengdu and Jinan Mandarin tones in the pre- and post-exposure conditions in the with-minimal-pair design. The results showed that native speakers of Standard Mandarin were able to understand sentences spoken in both unfamiliar dialects and successfully adapted to the new tone systems even before the explicit exposure. In general, explicitly presented passage exposure led to greater improvement for Chengdu sentences than Jinan sentences in both accuracy and response time.

In terms of accuracy, responses regarding sentence plausibility were significantly biased towards a plausible judgment in both dialects. Listeners considered the sentences in both Chengdu and Jinan Mandarin as plausible sentences regardless of the surprisal tone manipulation, which suggested listeners' high intelligibility of the unfamiliar dialectal utterances, but overall subpar identification of the mismatched tones. Similar to the findings in Chapter 3, they primarily relied on the available context and the familiar segmental information to understand the sentence meaning. This observed top-down mechanism for lexical processing was present in both Chengdu and Jinan speech and therefore, not tied to a particular dialect.

For the effect of dialect, listeners were generally more accurate in identifying mismatched tones in the Chengdu sentences than the Jinan sentences. It seemed that the more different Chengdu tone system was easier to adapt to than the more similar Jinan tone system, which confirmed the previous assumption: the phonetically more distinct sounds are easier to discriminate than the less distinct sounds relative to the listeners' native sound system. Moreover, listeners showed greater post-exposure improvement for the Chengdu sentences than for the Jinan sentences. Accuracy also gradually improved over trials after the explicit exposure in both dialects.

For response times, the credible effect of surprisal provided strong evidence of successful adaptation to the unfamiliar tone systems: listeners reliably slowed down for the surprisal tone across dialect and exposure conditions. The null credible effect of either dialect, exposure or trial on response time indicated listeners' early-on sensitivity to the high-surprisal manipulation in both dialects, with and without explicit exposure; incidental exposure to the sentence stimuli over the course of experiment was indeed sufficient to initiate adaptation, as suggested in the previous experiments in Chapter 3. Crucially, listeners seemed rapidly adapting to the unfamiliar tones as soon as they received the initial input.

For between-dialect differences, listeners further slowed down for the Chengdu sentences after receiving passage exposure than for the Jinan sentences after exposure. Specifically, they were slower in judging whether the Chengdu sentences sounded semantically plausible after the exposure, but faster in judging the Jinan sentences after the exposure. It seemed that the explicit exposure to the unfamiliar tone system improved listeners' discrimination of the surprisal tones in Chengdu Mandarin, but not quite so in Jinan Mandarin. Furthermore, faster responses were found over trials after exposure in the Chengdu condition than the Jinan condition as listeners made faster responses for the Chengdu low-surprisal sentences after the exposure.

For both accuracy and response time, the effect of explicit passage exposure varied between the Chengdu speech and the Jinan speech. It is likely that the phonetically more dissimilar tones attracted more attention during the explicit exposure and listeners grew more sensitive to the surprisal tones in Chengdu Mandarin than in Jinan Mandarin. Accuracy and response time results both indicated better adaptation to the less similar tone system, i.e. the Chengdu Mandarin tone system.

The major findings in Experiment 3 confirmed those in Experiment 1 and 2 in the previous chapter: credibly lower accuracy in the high-surprisal conditions and consistent slowdown for the high-surprisal sentences. It was reassuring to found similar patterns for Jinan Mandarin, so the conclusions we drew from these experiments can generalise to adaptation patterns in general, rather than due to a specific tone inventory. As for the experimental design, the previous findings suggested that minimal-pair presentation is not necessary for adaptation to occur for Chengdu Mandarin (Experiment 2); we wonder whether this is true for Jinan Mandarin as well. The current Experiment 3 used the with-minimal-pair design. Experiment 4 in the following section replicated the procedures of Experiment 3 with a no-minimal-pair design.

4.3 Experiment 4: Adaptation to Chengdu vs. Jinan tones with no minimal pairs

4.3.1 Methods

4.3.1.1 Participants

Another group of seventeen native speakers of Standard Mandarin participated in the experiment. Participants were randomly assigned to either the Chengdu task (9 participants) or the Jinan task (8 participants).

4.3.1.2 Materials

The same twenty-four sentence pairs were used in this experiment as in Experiment 3. In Experiment 4 we used a standard Latin square design with each sentence frame appearing only once in each surprisal and exposure condition. For each dialect, four Latin Square lists of sentences were created manipulating surprisal and exposure conditions; each list contained 24 unique sentence frames. Each sentence frame only appeared once in one of the four surprisal and exposure conditions (high-surprisal pre-exposure, low-surprisal pre-exposure, high-surprisal post-exposure, or low-surprisal post-exposure condition). The participants were assigned with either a Chengdu list or a Jinan list. The same passage recordings were used as in Experiment 3. Tone categories were counterbalanced across all the conditions.

4.3.1.3 Procedure

The procedures were identical to Experiment 3: the practice trials were followed by test trials, which started from the pre-exposure sentence semantic plausibility judgment task, then the exposure phase, and the post-exposure task in the end.

4.3.1.4 Data analysis

Accuracy and response time from the test phase were analysed as dependent variables across manipulations of surprisal (high-surprisal vs. low-surprisal), exposure (pre-exposure vs. post-exposure) and dialect (Chengdu Mandarin vs. Jinan Mandarin). Extreme values in response time outside the mean RT ±1 standard deviation range were excluded. There were 408 trials (24 trials × 17 participants) and 390 trials after the outlier exclusion.

The same models for accuracy and response time were used, except that tone category was added as an additional main factor and interacted with surprisal, exposure and dialect. Deviation coding was used to compare the mean of the dependent variable for phonological Tone 1, Tone 2, Tone 3 respectively to the overall mean of the dependent variable. Tone 4 was not compared to all the levels.

4.3.2 Results

4.3.2.1 Accuracy (Experiment 4)

As in Figure 4.4, accuracy for the low-surprisal sentences was considerably higher than the high-surprisal sentences, but the average accuracy value of the low-surprisal conditions (low-surprisal mean accuracy = 80.1%) was lower compared to that in Experiment 3 (low-surprisal mean accuracy = 91.3%). It was particularly low for Jinan Mandarin conditions (Jinan low-surprisal mean accuracy = 76.8%) compared to 89.2% in the previous experiment. Accuracy did not necessarily increase after the explicit exposure as in Experiment 3. In the Chengdu conditions, accuracy dropped for both low- and high-surprisal sentences after exposure, while it was higher for Jinan sentences after exposure.

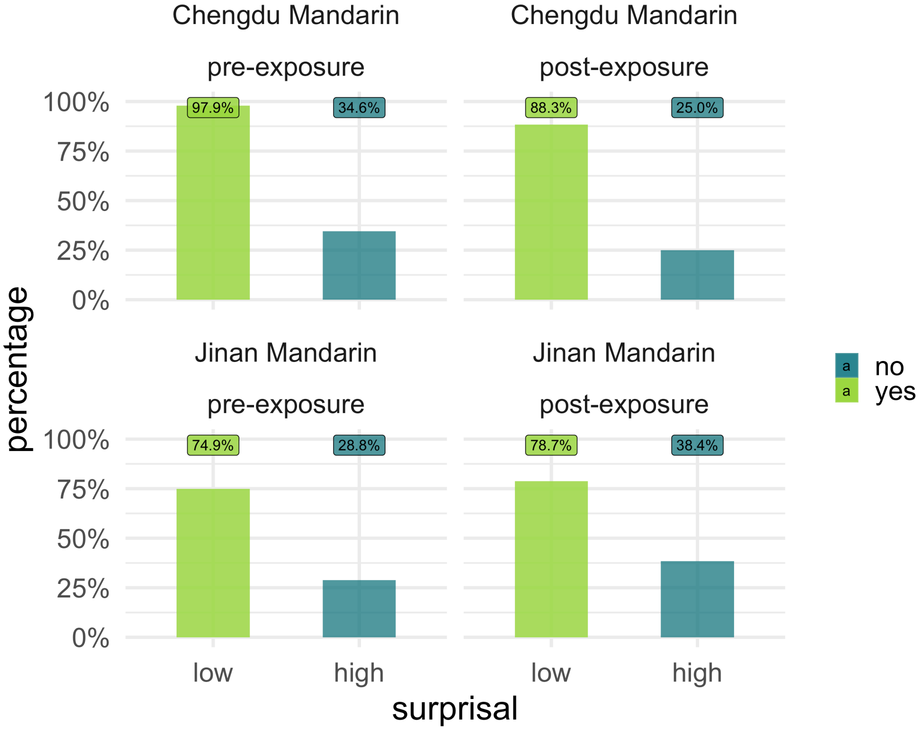


Figure 4.4 Percentage of "correct" responses across dialect, exposure and surprisal conditions. "Yes" (plausible; lime green) is treated as correct for low-surprisal conditions, and "no" (not plausible; blue green) for high-surprisal conditions.

The model revealed credible main effects of surprisal, the interaction between dialect and exposure, and the interaction between dialect and Tone 1. Specifically, accuracy was reliably higher in the low-surprisal condition than in the high-surprisal condition (surprisal: β = −2.77, 95% CI = [−4.33, −1.52]). A credible interaction was observed between dialect and exposure, but not in the expected direction, indicating unexpected lower accuracy for sentences spoken in the Chengdu dialect after exposure (dialect x exposure: β = −0.53, 95% CI = [−0.97, −0.14]).

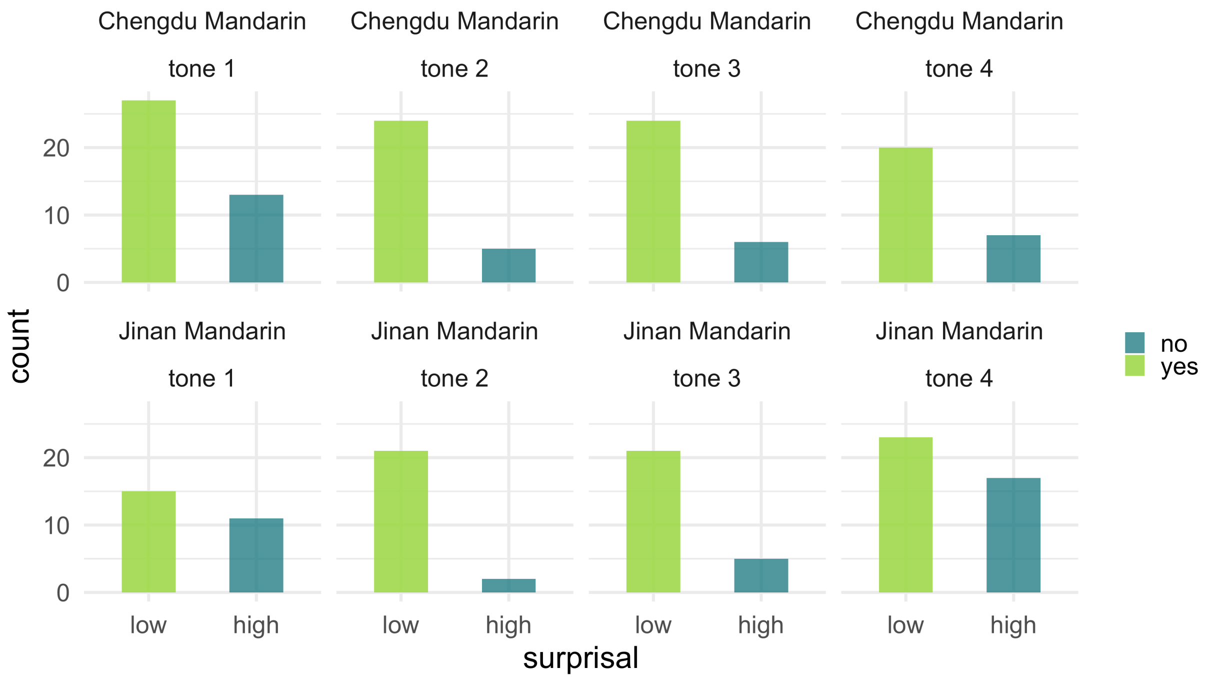


Figure 4.5 Count of "correct" responses across dialect and surprisal conditions by tone categories.

The model also detected a credible interaction between dialect and Tone 1, suggesting higher accuracy for sentences with Tone 1 manipulation in the Chengdu condition than the Jinan condition (*dialect x Tone 1:* β = 0.90, 95% CI = [0.22, 1.70]). No credible effect was found for the interactions of dialect with the other tone categories (*dialect x Tone 2*: β = 0.48, 95% CI = [−0.20, 1.15]; *dialect x Tone 3*: β = 0.24, 95% CI = [−0.38, 0.88]). Though Tone 4 was not compared given the coding scheme, accuracy was considerably higher for Jinan Tone 4 relative to the average, as in Figure 4.5.

4.3.2.2 Response time (Experiment 4)

The distributions of participant-specific response times for each condition are presented in Figure 4.6. Surprisingly, no reliable main effects were observed for dialect, surprisal, exposure, trial, tone category, or any of the interactions (dialect: β = 34.73, 95% CI = [−236.47, 312.06]; surprisal: β = 45.64, 95% CI = [−77.05, 168.74]; exposure: β = −2.88, 95% CI = [−77.16, 72.34]; trial: β = −0.81, 95% CI = [−45.63, 41.70]; dialect x Tone 1: β = −62.02, 95% CI = [−221.07, 89.89]; dialect x Tone 2: β = 12.31, 95% CI = [−126.18, 153.68]; dialect x Tone 3: β = 13.53, 95% CI = [−120.32, 156.68]. These seemed to suggest that with the no-minimal-pair design, participants might not be aware of the high-surprisal manipulation, and adaptation was not statistically credible before and after exposure. It is possible that the tested factors in Experiment 4 were not statistically credible due to the relatively small data size.

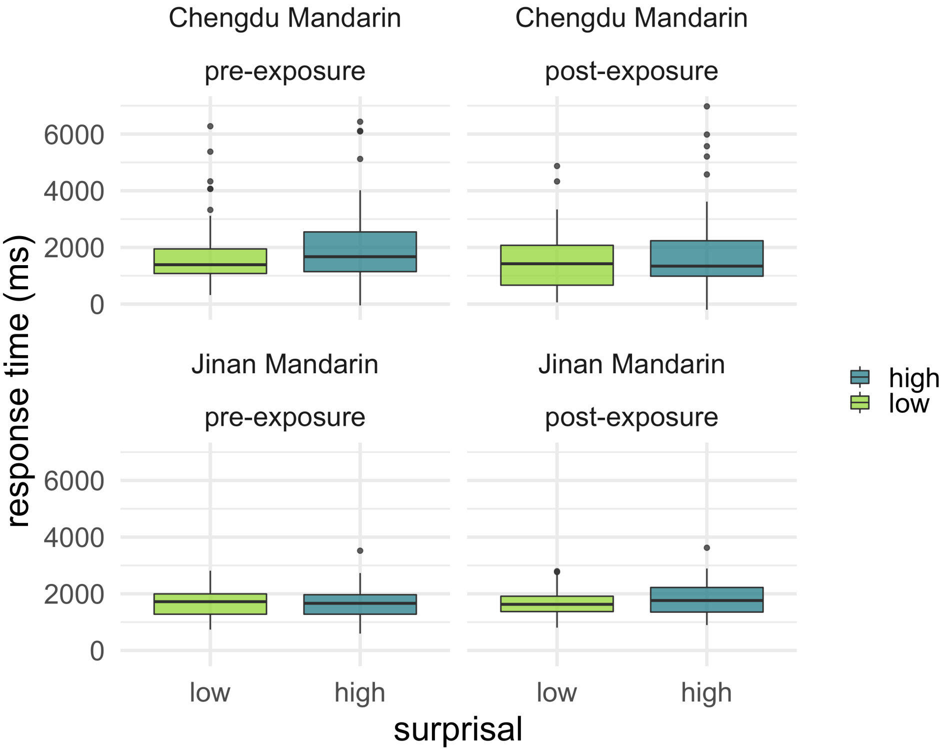


Figure 4.6 Response times across dialect, exposure and surprisal conditions.

Though the model revealed no reliable difference in response times by tone categories, numerically speaking, Tone 1 seemed to attract more attention in high-surprisal condition in both dialects (Figure 4.7), which might indicate possible better discrimination of Tone 1 manipulation. Also, participants seemed particularly good at responding to Tone 4 manipulation in Jinan Mandarin.

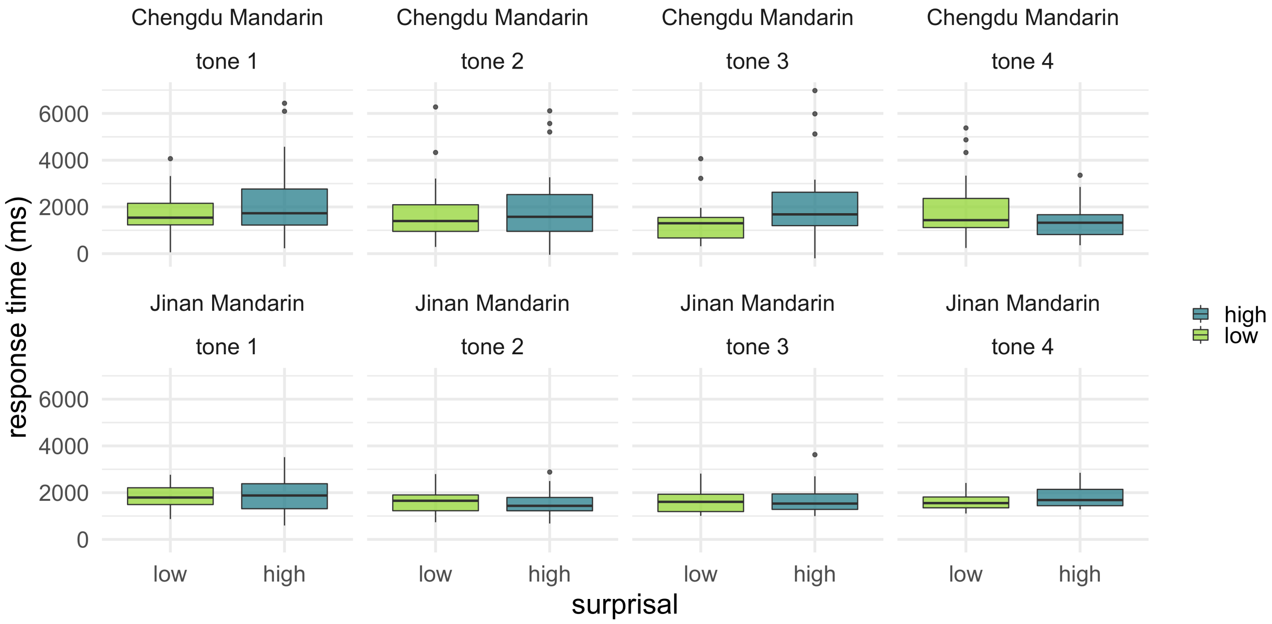


Figure 4.7 Response times across dialect and surprisal conditions by tone categories.

4.3.3 Discussion

In Experiment 4, we removed the minimal-pair presentation of the stimuli by using a Latin Square design. The results showed that listeners had no difficulty understanding the dialectal sentences without the minimal-pair contrast; however, response time results did not reveal difference between high and low surprisal, suggesting no adaptation to the unfamiliar tones in Chengdu and Jinan Mandarin.

Accordingly, listeners had similar low accuracy for the high-surprisal sentences as in the previous experiments. They were able to understand the sentences in both Chengdu and Jinan Mandarin but failed to report the mismatched tone in the high-surprisal condition.

However, no between-dialect difference was found on accuracy in this no-minimal-pair experiment. It seemed that listeners were equally bad at making the plausibility judgment for the high-surprisal sentences in both dialects. Moreover, there was no credible effect of exposure or its interactions with the other factors on accuracy for both dialects.

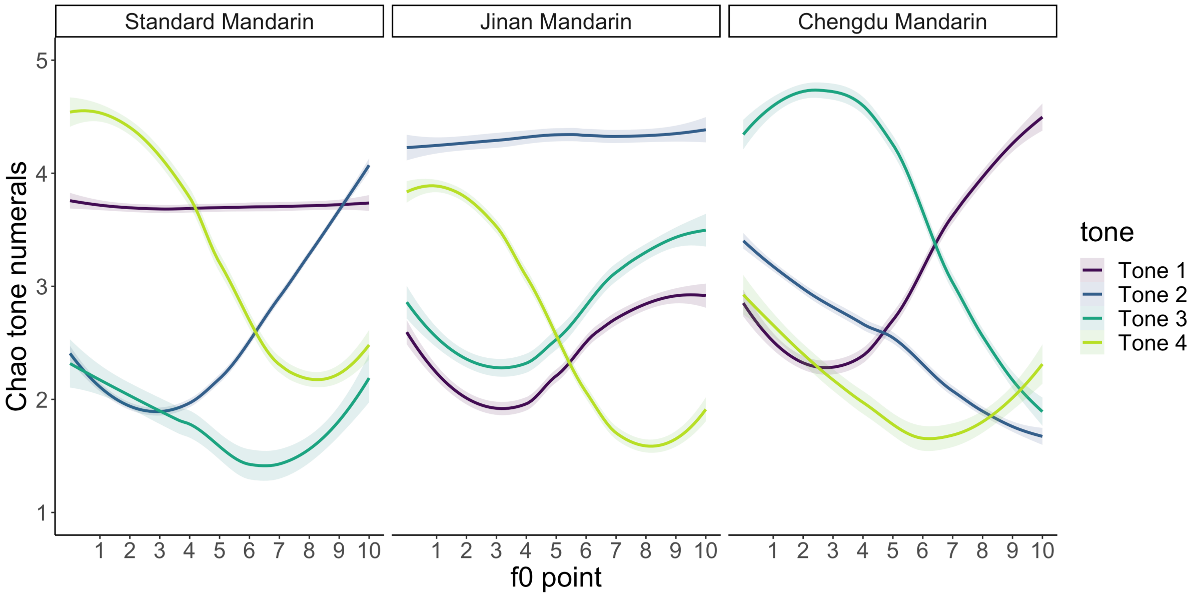


Figure 4.1 (repeated) Smoothed lexical tone contours of Standard, Jinan and Chengdu Mandarin converted to Chao Tone numerals (Zhao & Chodroff, 2022). Ribbons reflect ± 0.2 standard error of the mean.

For tone-specific adaptation, listeners were generally good at processing Chengdu Tone 1 and presumably Jinan Tone 4 relative the average and achieved higher accuracy for sentences containing the target words with these two tones. First, let us keep in mind that Tone 1 in Standard Mandarin is a level tone (see Figure 4.1 repeated here).

In Chengdu Mandarin, no level contour exists in the tone system and Chengdu Tone 1 is the only category that has a rising contour, therefore unique in terms of the contour type. When native Standard Mandarin listeners received the dialectal speech input, it might be the case that they became aware of the possible phonetic contours allowed in this unfamiliar tone system. Upon hearing Chengdu sentences, they might start to have a vague impression that there should be a rising contour of some unsure phonological category, and some falling contours, as well as a dipping contour with a long falling portion. As the input accumulated, top-down contextual information would help disambiguate the phonological category of the heard contour. For Chengdu Tone 1 specifically, when the sentential context strongly predicted the target word as the one of Tone 1 category, listeners could then map the unfamiliar contour onto the Tone 1 category. Crucially, the remapping of tone category with the novel phonetic contour might be easier if the unfamiliar contour is unique in its contour type as there are no other similar contours to compete with and therefore complicate the process.

For Jinan Tone 4, the remapping seemed rather smooth since Jinan Tone 4 has the same phonetic contour as the native Tone 4, both a falling contour. It was likely that the words of Tone 4 category in Jinan Mandarin sounded native-like for listeners of Standard Mandarin. Therefore, the accuracy results of Tone 4 in Jinan Mandarin somewhat resemble the patterns found in the native dialect (see Figure 3.4 in the Standard Mandarin conditions) – relatively high accuracy in both surprisal conditions. It seems that the tone with the same contour–category mapping as the native one is easier to adapt to compared to those with different mapping schemes.

With respect to response times, the model detected no credible effect of any fixed factor or interaction, while the previous experiments all found a reliable difference in response times between low and high surprisal. We suspected that the differing results might be attributed to the Latin Square design in Experiment 4, or the reduced number of trials before and after exposure, or merely the insufficient data from fewer number of participants and potential between-group perceptual differences.

Both Experiment 2 (Section 3.3) and Experiment 4 implemented the no-minimal-pair design, but differed in the structure of the sentence stimuli. Specifically, in Experiment 2, the minimal-pair presentation was removed by assigning each surprisal version of the sentence to a different dialect and listeners heard all the Chengdu sentences before all the Standard Mandarin sentences. In this sense, listeners heard the unique sentence frames for the first time in the unfamiliar Chengdu Mandarin and then the same set of frames in different surprisal versions in their native speech. There were twenty-four trials in the Chengdu condition in Experiment 2; thirteen listeners participated in the Chengdu task. In Experiment 4, the Latin Square listing was applied to ensure each sentence frame was presented once across the surprisal and exposure conditions. In each dialect, listeners heard twelve unique items before the explicit exposure and the other twelve unique items after the exposure. There were nine participants in the Chengdu task and eight participants in the Jinan task.

Although the two experiments varied in the structure of the stimuli, listeners never received a sentence frame twice in the unfamiliar speech. It seems unlikely that it was the Latin Square design that obviated the adaptation process. However, by the Latin Square design, listeners heard half of the sentence frames in the pre-exposure condition and half in the post-exposure condition, which reduced the amount of sentential input in each exposure condition. Although by the end of the experiment, listeners received twenty-four sentence trials as in Experiment 2, passage exposure was inserted in the middle of the stimuli, which might potentially impede the adaptation process. It was possible that listeners had to switch to a different type of input information using different learning mechanisms before newly formed or updated representations could consolidate over the reduced sentence trials.

Also, in the following section 4.4, we provided the combined analysis of the effect of dialect and exposure with the data from both experiments.

4.4 Combined analysis

Based on the findings in Chapter 3, we expected similar adaptation patterns between Experiment 3 and 4 since minimal-pair presentation was found unnecessary to initiate adaptation. However, a great deal of null effect was found in Experiment 4 for the factors which were expected to be credible given the previous results. As both Experiment 3 and 4 had a relatively small sample size, it was beneficial to run a combined analysis which could improve our estimates of the effect of surprisal, dialect and exposure.

4.4.1 Data analysis

We combined the data from Experiment 3 and 4 in assessing the effect of explicit exposure, dialect, and surprisal on accuracy and response time in the sentence plausibility judgment task. The accuracy and response time models were the same as for Experiment 4, except that for the accuracy time model, the experimental design was added as a main effect and sum coded as with-minimal-pair design = 1, no-minimal-pair design = -1, and for the response time model, experimental design (*design*: with-minimal-pair design = 1, no-minimal-pair design = -1) was added with the full interactions with dialect and surprisal. The interactions of design were not included in the accuracy model as they were not indicative of tone adaptation.

4.4.2 Results

4.4.2.1 Accuracy

Percentage accuracy results are shown in Figure 4.8. According to the model, credible main effects emerged for surprisal and exposure, but not for any interactions. For the surprisal effect, accuracy in the high-surprisal condition was reliably lower than that in the low-surprisal condition (*surprisal:* β = −1.43, 95% CI = [−1.66, −1.21]). For the effect of exposure, accuracy credibly improved after the explicit exposure (*exposure:* β = 0.29, 95% CI = [0.02, 0.59]). There was no credible interaction between dialect, surprisal and exposure (*dialect x surprisal:* β = −0.05, 95% CI = [−0.27, 0.17]; *dialect x exposure:* β = 0.13, 95% CI = [−0.07, 0.35]; *surprisal x exposure:* β = 0.15, 95% CI = [−0.08, 0.37]; *dialect x surprisal x exposure:* β = 0.03, 95% CI = [−0.19, 0.25]). Also, no credible effect was found for the minimal-pair design (*design:* β = 0.00, 95% CI = [−0.28, 0.29]), indicating no credible effect of the minimal-pair presentation on accuracy. The effect of trial was not credible either (*trial:* β = −0.01, 95% CI = [−0.03, 0.02]).

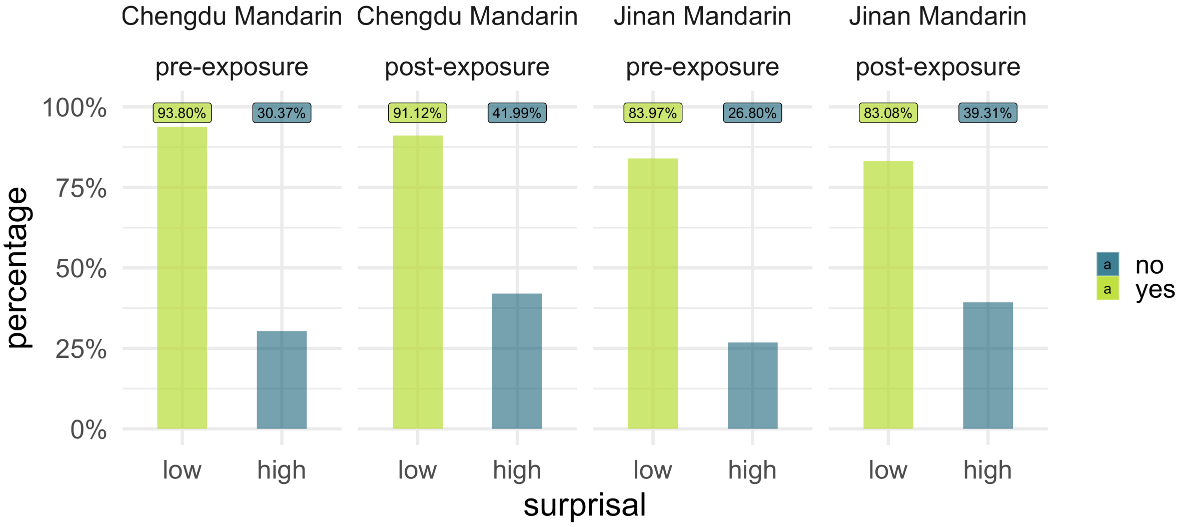


Figure 4.8 Percentage of "correct" responses across surprisal, dialect, and exposure conditions in the combined analysis.

4.4.2.2 Response time

Figure 4.9 presents the response times across surprisal, dialect and exposure conditions. The response time model detected credible main effects for surprisal, dialect, the interaction between dialect, surprisal and presentation, the interaction between surprisal and dialect, and the interaction of surprisal, dialect and exposure. Specifically, responses were reliably slower in the high-surprisal condition (*surprisal:* β = 0.07, 95% CI = [0.02, 0.11]). Listeners consistently slowed down for the high-surprisal sentences in both dialects, indicating successful adaptation to the unfamiliar tone systems. For the between-dialect difference, response times were credibly faster in the Chengdu Mandarin condition than Jinan Mandarin (*dialect:* β = −0.07, 95% CI = [−0.12, −0.02]. Response time was also credibly modulated by the interaction between surprisal and dialect (*surprisal x dialect:* β = 0.07, 95% CI = [0.03, 0.11]) and the interaction of surprisal, dialect and exposure (*surprisal x dialect x exposure:* β = 0.06, 95% CI = [0.01, 0.10]). The interaction results showed that the slowdown in the high-surprisal condition was greater for the Chengdu sentences than the Jinan sentences. Furthermore, responses were even slower after the passage exposure in the Chengdu high-surprisal condition, compared to the Jinan high-surprisal condition.

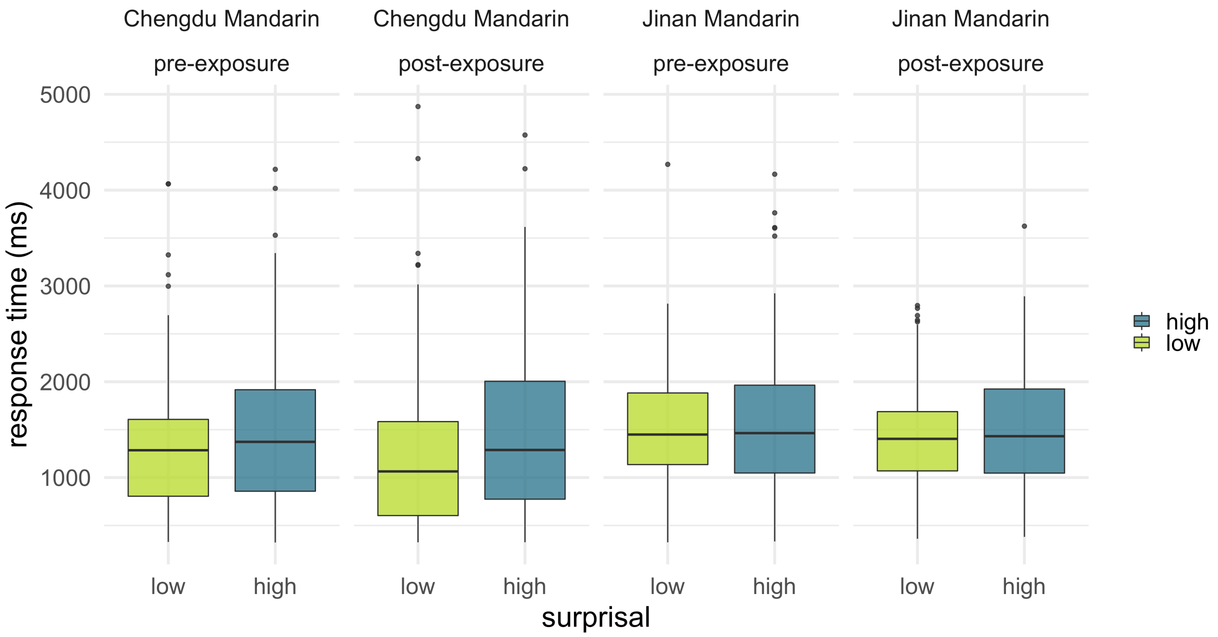


Figure 4.9 Response times across dialect, exposure and surprisal conditions in the combined analysis.

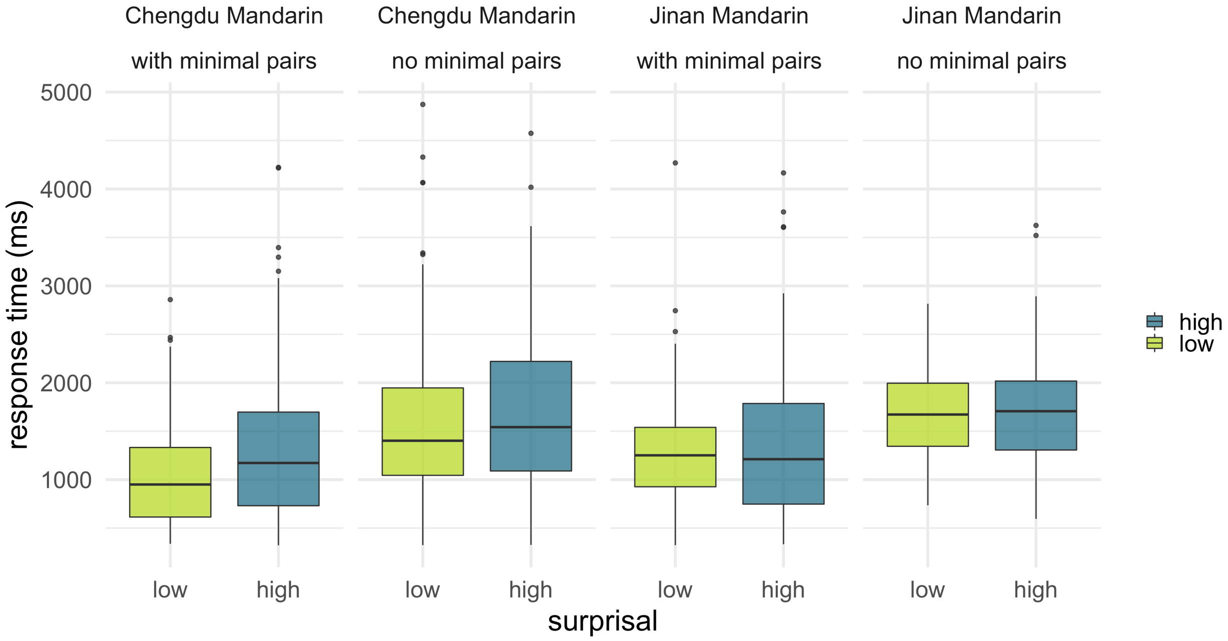


Figure 4.10 Response times across dialect, surprisal and presentation conditions in the combined analysis.

For the effect of experimental design (Figure 4.10), responses were generally faster when the minimal-pair sentences were present in the stimuli for both dialect conditions (design*:* β = −0.20, 95% CI = [−0.27, −0.13]). This may suggest faster and easier adaptation process when listeners receive a tonal contrast in the stimuli. However, since for each design, a different group of participants were tested; it is equally possible that the between-design difference was due to group-level differences. One group of participants might be generally slower or faster in making the plausibility judgment. In addition, experimental design did not credibly interact with either dialect, surprisal, or both, suggesting that minimal-pair-presentation did not credibly modulate response times across surprisal and dialect conditions, which conformed to the findings in Chapter 3. No credible effect of trial was found in the model (*trial:* β = 0.00, 95% CI = [−0.02, 0.02]).

4.4.3 Discussion

With the combined data, we particularly focused the effects of exposure and dialect on accuracy and response time in the sentence plausibility judgment task. For accuracy, listeners' identification of the surprisal tone improved after about two minutes of explicit exposure for both dialects and there was no between-dialect difference in accuracy. With the increased amount of word–tone information from the explicit exposure, together with more incidental exposure in the post-exposure trials, listeners reported the high-surprisal tones more accurately, especially in the high-surprisal condition. In addition, the lack of credible interactions between surprisal, dialect and exposure on accuracy suggested that post-exposure improvement in the high-surprisal condition was statistically equal in Chengdu and Jinan Mandarin. It seemed that dialectal differences had little impact on the accuracy of plausibility judgment. It is likely that regardless of whether listeners adapted to the new tone system or not, they primarily relied on top-down information in lexical decision; and this approach was applied in each dialect condition.

For response time, although explicit exposure did not credibly affect response times in general, listeners were more sensitive to the high-surprisal manipulation after the explicit exposure in the Chengdu condition than the Jinan condition. In fact, they slowed down more for the high-surprisal sentences in Chengdu Mandarin than Jinan Mandarin regardless of the explicit exposure. These results strongly indicated listeners' better adaptation to the unfamiliar tones in Chengdu Mandarin than Jinan Mandarin. It seemed that phonetically less similar tones were easier to adapt to with and without the explicit exposure. Passage exposure was more likely to function as a facilitating factor in boosting the discrimination of the surprisal tones, with the facilitation effect greater in the dissimilar tone system (Chengdu Mandarin) than the more similar tone system (Jinan Mandarin).

4.5 Conclusion

In this chapter, we looked into the effect of an explicit exposure period, as well as dialect on listeners' processing of unfamiliar tone systems. We found that listeners processed the unfamiliar bottom-up tone information when hearing unfamiliar dialects while primarily using top-down information to guide sentence plausibility judgment. Better adaptation was found for the dialect with a more dissimilar tone system to the listeners' native tone system, particularly when the tonal contrast was present in the stimuli. Explicit passage exposure helped in improving accuracy in lexical decision. It seems that the effect of passage exposure on the adaptation to unfamiliar dialect was more salient for the tone system that is phonetically less similar to the listeners' native tone inventory. The theoretical implications of these findings will be further discussed in the Chapter 5.

4.6 Authorship and publication status

The current chapter was written in preparation for a journal article, which has some self-contained components that may overlap with the other chapters in the thesis. The portions of the content were previously presented at two conference venues: “Rapid adaptation to unfamiliar Mandarin dialect tone systems: Evidence from bottom-up tone processing”, a talk at the Colloquium of the British Association of Academic Phoneticians (BAAP 2022); and “Rapid adaptation to unfamiliar lexical tone systems: the effects of dialect and explicit exposure” in the oral session at the Phonetics and Phonology in Europe Conference (PaPE 2023). The abstracts and slides were drafted by Liang and revised by Eleanor. The journal paper will be co-authored by Liang Zhao, Shayne Sloggett, and Eleanor Chodroff. The relative contribution is expected as follows:

Liang Zhao: Conceptualisation, methodology, investigation, resources, data curation, formal analysis, writing—original draft, writing—review & editing, visualisation; Shayne Sloggett: methodology, formal analysis, writing—review & editing; Eleanor Chodroff: methodology, visualisation, formal analysis, writing—review & editing, supervision, funding acquisition.

Chapter 5 Conclusion

The last chapter in the dissertation first summarises the key findings from the production and perception studies in the previous research chapters in Section 5.1. Linking back to the beginning of Chapter 1, Section 5.2 overviews the case of perceiving dialectal Mandarin speech with unfamiliar tone systems and discusses the theoretical implications of the findings, as well as a minor aspect in the experiments for follow-up investigation. Section 5.3 states the limitations of the current study which point to future directions of research.

5.1 Summary of the key findings

5.1.1 The production study: comparable segments and distinct phonetic tones

The production study in the dissertation comprised the acoustic-phonetic analysis of the vowel inventories and the lexical tone systems of six Mandarin dialects—Beijing Mandarin, Chengdu Mandarin, Jinan Mandarin, Taiyuan Mandarin, Wuhan Mandarin and Xi'an Mandarin (Chapter 2). The study provides an update on the Chao tone numerals for each lexical tone category in each of the six dialects, such that they more closely resemble the actual phonetic realisation in natural speech compared to older descriptions of these systems. According to the measured acoustics and results from the statistical models, the production study establishes that Mandarin dialects have comparable segmental systems and distinct phonetic tone inventories. For differences between the Mandarin tone systems, there are two types of variation in general: the tone systems that are phonologically different and those that are phonetically different, but phonologically consistent. Among the six Mandarin dialects, Taiyuan Mandarin differs from Standard Mandarin in the number of phonological tone categories; the other dialects all have a four-tone system and have relatively consistent word–

tone mappings relative to Standard Mandarin. However, these phonologically similar tone systems differ considerably in their phonetic inventories. Some of these dialects have less similar phonetic contours compared to Standard Mandarin (e.g. Chengdu Mandarin); some have similar phonetic contours, just that the word–tone mappings are different (e.g. Jinan Mandarin).

Contributing to the findings on lexical tone variation is a methodological project for remote audio collection and speech data annotations, published as the ManDi corpus. The current version of the corpus contains 357 recordings (about 9.6 hours) of monosyllabic words, disyllabic words, short sentences, *the North Wind and Sun* passage and a Chinese anecdotic poem, each produced in Standard Mandarin and in one of six regional Mandarin dialects: Beijing, Chengdu, Jinan, Taiyuan, Wuhan, and Xi'an Mandarin from 36 speakers.

5.1.2 The perception studies: rapid adaptation through integrative perception

The most surprisal finding in the perception studies is the rapid adaptation to the unfamiliar tone systems under adverse conditions. Listeners are able to adapt to the novel phonetic tones with about one minute of incidental exposure and without explicitly presented tonal contrasts. Moreover, the statistical analysis indicates that such adaptation may occur as early as the experiment commences. These findings are consistent with Clarke and Garrett's (2004) study. Both the current study and Clarke and Garrett's (2004) suggested that perceptual adaptation to unfamiliar phonetic variation occurs after about one minute of sentential exposure and this process may start early on upon the initial input.

During this highly efficient adaptation process, bottom-up and top-down mechanisms are jointly integrated in processing the lexical tone variation. For the familiar tone system (Standard Mandarin), listeners use both top-down and bottom-up information, as both sources are reliable for native listeners. For the unfamiliar tone systems (Chengdu Mandarin and Jinan Mandarin), listeners resort to top-down information in making the lexical decision, but they also actively process the novel tone acoustics in a bottom-up manner and readily update the tone–contour mappings in the unfamiliar tone system. Therefore, we conclude that phonetic variation in lexical tone systems can be processed with integrated top-down and bottom-up mechanisms for both familiar and unfamiliar tone systems. There is an overriding top-down influence on deciding the word identity that is congruent with the prior context, while the bottom-up processing of the novel tone acoustics commences as soon as the speech signal becomes available. These findings confirmed the assumption of constructive speech perception (Section 1.3), where listeners make use of all available information from the input for lexical access, especially when the bottom-up processing leads to ambiguous or unfamiliar lower-level output.

The perception studies also tested several factors concerning the quality and quantity of the exposure and dialectal difference between tone systems in adaptation to lexical tone variation. Firstly, minimal-pair tonal contrasts in the sentence stimuli are not necessary for the adaptation to occur. Listeners may use lexical contrast to consolidate newly constructed tone-contour mappings, but absence of the tonal contrast does not obviate the adaptation process. Similar to Hayes-Harb's (2007) finding on segmental adaptation, the indispensable information in the input is the well-presented systematic patterns, rather than specific lexical contrasts, such as minimal pairs.

Secondly, a minimum amount of incidental exposure is sufficient to induce adaptation, while both the increased input from more sentential trials and the explicit passage exposure can facilitate adaptation with more accurate lexical decision, which is consistent with previous findings on both adaptation to novel phonetic contrast (Norris et al., 2003) and adaptation to accented speech (Bradlow & Bent, 2008).

Thirdly, perception of the lexical tone variation is modulated by the relative similarity of the tone contours compared to the listeners' native tone system. In line with the Perceptual Assimilation Model (So & Best, 2010; Reid et al., 2015), the tone system with more dissimilar contours relative to the native system is easier to adapt to in terms of response times in lexical decision. Additionally, greater improvement after the explicit exposure is found for dissimilar tones (Chengdu tones) than the more similar tones (Jinan tones).

The major findings on the adaptation to dialectal Mandarin speech add to the current understanding of speech adaptation, particularly on adaptation to phonetic variation on the suprasegmental level. The perception studies in this dissertation showed that previous hypotheses on the influential factors and the expected effects can generalise well to adaptation to lexical tone variation.

5.2 Further discussions

Mandarin dialects vary substantially in the phonetic inventory and particularly in the phonetic realisation of lexical tones. To return to the conundrum between the word "*jiao1 dai5*" (an explanation/a solution) and "*jiao1 dai4*" (a roll of tape), when hearing the word in an unfamiliar dialect, such as Chengdu and Jinan Mandarin, what the non-native listener perceive from the acoustic information are the native-like segments and some unexpected tone contours. If the word is presented in isolation, misunderstanding might still be there since the acoustic information alone cannot disambiguate between the two options; moreover, the tone acoustics itself is unfamiliar.

However, in real-life communication, words are most often recognised with other sources of information alongside of the acoustics. Suppose a conversation happens when a Chengdu Mandarin-speaking neighbour is up in the tree house to do some repairing and accidentally drops the hammer and the tape onto the ground; he/she might ask someone passing by and say "hi, could you help me pick up the hammer and the *jiao1 dai4?*" in Chengdu Mandarin. The passer-by does not necessarily speak Chengdu Mandarin natively, but as long as he/she has knowledge of any Mandarin language, the word can readily be accessed as a roll of tape. They might even engage in a few minutes' small talk and both find each other's dialectal speech sounding more intelligible. (This example may involve extralinguistic cues for word identification; the discussion here is limited to linguistic cues only.)

5.2.1 The presence of top-down influence in lexical access

In the above example, word recognition is dominated by the preceding context together with the familiar segmental information, rather than the novel tone acoustics, as the listener has no idea of how these familiar-sounding syllables (i.e. "*jiao*" and "*dai*") should be realised phonetically in the unfamiliar dialect. In accessing word meaning, it is likely that top-down expectedness of the word identity overwrites the unfamiliar acoustic information; the lexical meaning is therefore extracted under the influence of higher-level processing.

At this step of reasoning, some might question whether it is possible that the listener has already learned the new tone–contour mappings in the unfamiliar dialect, so the word identification is still a bottom-up process. This question can hardly be answered if the acoustic information is all legitimate in the input because either bottom-up processing or top-down processing would output the same lexical meaning—"adhesive tape". Certain discrepancy between the bottom-up and top-down information needs to be created so that the listener' lexical decision can be indicative of which mechanism takes effect.

In the perception studies of this dissertation, the necessary discrepancy to deconfound expectation from acoustics was created using the surprisal sentence pairs. A target word in each sentence item was altered in terms of lexical tone category to create a semantically implausible version of the sentence, i.e. the high-surprisal sentence, in contrast with the default semantically plausible version without tone manipulation, i.e. the low-surprisal sentence. The listener's task is to judge the semantic plausibility of the sentence which entails the lexical decision of the target word. If listeners indeed manage to quickly learn the new tone–contour mappings in the unfamiliar tone system, they would identify the high-surprisal tone and report the sentence as implausible. However, the results in the perception studies suggested otherwise—listeners reported the high-surprisal implausible sentences mostly as plausible in the unfamiliar speech; this pattern was statistically credible in all the conditions of the perception experiments (Experiment 1, 2, 3 and 4).

The ultimate bias towards a plausible judgment even though the bottom-up tone acoustics suggests implausibility provides concrete evidence for the presence of top-down influence in lexical access, empirically tested on lexical tone processing in a tonal language. Our finding conforms to the assumptions in the interactive-activation mechanism of the TRACE model (Elman & McClelland, 1984; McClelland & Elman, 1986) and the hypothesis of top-down synthesis for word recognition in the model of acoustic landmarks and distinctive features (Stevens, 2008). It is in stark contrast to the Shortlist model (Norris, 1994) and the Merge model (Norris, 1994; Norris, McQueen & Culter, 2000), which argued for an entire bottom-up and denies the necessity of top-down feedback in guiding online processing. They considered the top-down influence as offline feedback for learning which functions as a computable "error-correcting signal" (Norris et al., 2003), but our finding suggested that when the acoustic information is unfamiliar and less reliable, bottom-up processing can be affected, or even overridden by information from higher-level processing.

5.2.2 The relative contribution of bottom-up and top-down information: modifiable given reliability

For the perception of lexical tone variation, the perception studies suggested an integrated use of bottom-up and top-down mechanisms for both familiar and unfamiliar speech. With different sources of information at hand, the perceptual system seems exceptionally good at actively modifying their relative weight to achieve *maximum efficiency with minimum effort* in understanding the intended meaning. To be specific, the relative contribution of top-down and bottom-up information may vary given the relative reliability of the information. When bottom-up information reliably estimates a matching candidate, it is likely that the effect of top-down processing is deemed unnecessary and minimised accordingly. Likewise, when top-down information provides strong lexical expectedness, while acoustic information is insufficient or results in recognition of unfamiliar speech sounds inconsistent with the context, bottom-up information can be entirely renounced in the lexical decision.

This assumption of modifiable processing weight for different mechanisms can nicely explain the case with perception of lexical tone variation. For the unfamiliar phonetic tones, when the sentential context strongly predicates a lexical item, but the bottom-up tone acoustics is new and less reliable, listeners quickly opt for a top-down solution in determining the lexical word. For the familiar tones, since both sources of information can be reliable, their relative contribution may depend on some external factors, such as the specific task, consistency of the context, or the stage at which the responses are measured.

For example, if the task is to report the lexical item following the highly predictive context in listeners' native sound system, they may rely on the bottom-up information alone since it is highly reliable. Even if the context is neutral or inconsistent with the target item, the identification can still rely on the bottom-up information because the task is to identify the target lexical item in listeners' native speech: they may be aware of the inconsistency, but this might not slowdown lexical decision if they "choose" to minimise the unnecessary processing effort. As a result, listeners can have similar response times in different context conditions. Actually, this is what the Connine's (1987) study found. They argued against the top-down context influence given the lack of difference in response time, but it might just be the case that listeners are savvy users of available information and automatically minimise the top-down influence as soon as they recognise that the heard speech is all native and bottom-up processing is good enough for this specific task.

Given the perceptual patterns found in the present perception studies, it is hypothesised that the relative contribution of bottom-up and top-down information can be modified given their relative reliability between the perception of familiar and unfamiliar speech. Similar notions can be found in Goldstone's (1998) proposal of "attention-weighting" for different stimuli and in the TRACE model's assumption of modifiable weighted connection between units of processing. The model assumes that the degree of activation allowed to flow between units on different levels of processing is strengthened through excitatory interactions between consistent units and diminished through inhibitory interactions between inconsistent units (Elman & McClelland, 1984; McClelland & Elman, 1986). The current dissertation does not have precise conclusions on this topic; it is worth investigating in future study.

5.2.3 Perceptual adaptation as remapping between tone category and phonetic contour

Spoken speech contains rich phonetic variation in the speech signal. Instead of treating such variation as a nuisance or an impediment to communication, the perceptual system tends to constantly update the mappings on different levels of representation to accommodate the novel input. From the above discussion, bottom-up processing seems inferior in terms of lexical decision in the unfamiliar speech. However, this does not mean that bottom-up information is not processed at all. In fact, the response time results in the perception studies uniformly indicated the presence of bottom-up processing of the unfamiliar tones, which is evidenced by the consistent slowdown for the high-surprisal sentences in both unfamiliar dialects, i.e. Chengdu Mandarin and Jinan Mandarin.

Listeners successfully adapt to an unfamiliar tone system through remapping between the phonological tone category and the phonetic contour. Therefore, when a high-surprisal tone is heard, they become aware of the wrong tone-contour mapping and direct more attention to the inconsistency between the context and the mismatched tone contour. However, the updated tone-contour mappings are not explicitly used in lexical decision; listeners still report the high-surprisal sentences as plausible, which leads to low accuracy in the sentence semantic-plausibility judgment. This pattern is consistent with the previous hypothesis that perceptual adaptation may not necessarily result in improvement in sound discrimination (Samuel & Kraljic, 2009). The possible reasons are elaborated in the above section.

5.2.4 The effect of increased exposure: potential ceiling effect?

The perception studies in the dissertation used the repetitions of the sentence trials (Experiment 2) and the explicit passage exposure (Experiment 4) to increase the amount of input of the unfamiliar speech. Both experiments showed that the accuracy was credibly higher after the increased exposure, be it incidental or explicit; sensitivity to the high-surprisal tone was also enhanced as the exposure increased in both dialects. A minor finding concerning the exposure effect might indicate a potential ceiling effect in how much improvement listeners can achieve with more amount of input.

Specifically in Experiment 2, listeners received sentential stimuli in Chengdu Mandarin over three repetition blocks. The results showed that accuracy in judging sentence plausibility reliably increased with another one-minute exposure in the second block. However, there was no credible improvement in accuracy from the second block to the third block. Although this is not tested in the current study, it is possible that the relation between the amount of exposure and the extent of accuracy improvement is logarithmic, rather than linear. Accordingly, the initial increase would trigger greater extent of improvement, but to observe the same extent of improvement in a later stage, a larger amount of exposure is needed. A follow-up study can be done to empirically investigate the relation between the amount of exposure and the adaptation outcome to see whether the improvement in lexical decision stagnates after certain amount of exposure.

5.3 Limitations and directions for future research

The perception studies in this dissertation investigated adaptation to the unfamiliar lexical tone systems with incidental and explicit exposure to the unfamiliar dialectal speech. The adaptation effect was examined based on the responses immediately following the stimulus signal. As we propose that impoverished representations might be constructed between the phonological category and the newly received tone acoustics, whether long-term representations are formed was not tested. Further research could be done on the memory-based perception to investigate whether adaptation to dialect-specific lexical tone variation persists after longer periods of time. The factors being tested in the perception studies are also worth assessing on the potential outcome of long-term learning of the novel tone system.

The tone-specific adaptation patterns were not adequately explored in the current study. Part of the reason is that for the manipulation of the main factors, tone categories were not perfectly balanced for each condition in each experiment. However, even with the tones nearly balanced, the results did not indicate clear patterns for the tonal factor. Though this is not the focus of this thesis as we aimed to explore adaptation to the unfamiliar tone system as a whole, rather than the individual tone categories, further study could be done to investigate category-specific adaptation by manipulating the between-dialect phonetic similarity more precisely using synthesised speech. The overall similarity of the phonetic tone inventory relative to the native tone system may also be quantified in this approach.

For cross-talker adaptation, the current experiments used the stimuli for each dialect produced by one speaker. The reason for the single-speaker exposure per dialect was to avoid attention directed to the across-talker difference, rather than similarity in speech itself. Future research could experiment on listeners' ability to generalise tonal patterns to novel speakers.

The current perception studies establishes an overriding effect of top-down information in processing unfamiliar lexical tones. The predictiveness of the top-down information is considered strong in the current design, but whether the degree of predictiveness of the sentential context would affect the adaptation outcome is unclear. In natural speech, it is not always the case that sentential context is consistent with the tone acoustics. What happens when the pre-target context contains errors as inconsistent cues? Can adaptation to the novel tones still occurs when the top-down information is not reliable? These questions could be tested in a lexical decision task or in an eye-tracking experiment, where participants identify the target word after hearing the sentence contrasting in top-down reliability.

More generally, all the experiments were conducted during the lock-down periods in China and in the UK, more data could be collected for the experiments, which hopefully could be achieved soon for the follow-up studies.

APPENDIX

The following lists contain the reading materials in the production experiment (Chapter 2) and the full set of surprisal sentences designed for Chapter 3 and 4. For the production data, the audio recordings, TextGrid annotations and scripts can be found in the open-access ManDi Corpus on OSF at https://osf.io/fgv4w/. The lists of sentence stimuli used in the perception experiments can be found at https://osf.io/gmc8e/.

a. List of monosyllabic words

Item	Character	Pinyin	Tone	English translation	Experimental phase
-	猜	cai	1	to guess	Familiarisation
-	才	cai	2	Just/merely	Familiarisation
-	彩	cai	3	colour	Familiarisation
-	菜	cai	4	vegetable	Familiarisation
1	八	ba	1	eight	Test
1	拔	ba	2	to pull out	Test
1	把	ba	3	to grasp	Test
1	爸	ba	4	dad	Test
2	夫	fu	1	husband	Test
2	服	fu	2	clothes	Test
2	腐	fu	3	decay	Test
2	复	fu	4	again	Test
3	遮	zhe	1	to cover	Test
3	哲	zhe	2	sage	Test
3	者	zhe	3	person	Test
3	这	zhe	4	this/these	Test
4	吃	chi	1	to eat	Test
4	池	chi	2	pool	Test
4	齿	chi	3	tooth/teeth	Test
4	赤	chi	4	red	Test
5	汤	tang	1	soup	Test

5	糖	tang	2	sugar	Test
5	躺	tang	3	to lie down	Test
5	烫	tang	4	hot	Test
6	懵	meng	1	confused	Test
6	萌	meng	2	cute	Test
6	猛	meng	3	fierce	Test
6	梦	meng	4	dream	Test
7	身	shen	1	body	Test
7	神	shen	2	spirit	Test
7	审	shen	3	to inspect	Test
7	慎	shen	4	cautious	Test
8	香	xiang	1	fragrant	Test
8	祥	xiang	2	auspicious	Test
8	想	xiang	3	to think	Test
8	向	xiang	4	towards	Test
9	优	you	1	excellent	Test
9	油	you	2	gas	Test
9	有	you	3	to have	Test
9	又	you	4	again	Test
10	一	yi	1	one	Test
10	姨	yi	2	aunt	Test
10	已	yi	3	already	Test
10	毅	yi	4	persistence	Test

b. List of disyllabic words

Item	Character	Pinyin	Tone	Translation	Experimental phase
-	住宅	zhu4 zhai2	4_2	housing	Familiarisation
-	牙齿	ya2 chi3	2_3	teeth	Familiarisation
-	吟唱	yin2 chang4	2_4	sing	Familiarisation
-	运气	yun4 qi5	4_4	luck	Familiarisation
1	呼吸	hu1 xi1	1_1	breath	Test
2	牵强	qian1 qiang2	1_2	forced	Test
3	优雅	you1 ya3	1_3	elegant	Test
4	风暴	feng1 bao4	1_4	storm	Test
5	知识	zhi1 shi5	1_5	knowledge	Test
6	房屋	fang2 wu1	2_1	house	Test
7	谣传	yao2 chuan2	2_2	rumour	Test
8	糖果	tang2 guo3	2_3	candy	Test
9	城市	cheng2 shi4	2_4	city	Test
10	麻烦	ma2 fan5	2_5	trouble	Test
11	普通	pu3 tong1	3_1	ordinary	Test
12	海洋	hai3 yang2	3_2	ocean	Test
13	打赌	da3 du3	3_3	bet	Test
14	巧妙	qiao3 miao4	3_4	ingenious	Test
15	尺子	chi3 zi5	3_5	ruler	Test
16	废墟	fei4 xu1	4_1	ruins	Test
17	送别	song4 bie2	4_2	farewell	Test
18	梦想	meng4 xiang3	4_3	dream	Test
19	探望	tan4 wang4	4_4	visit	Test
20	秘书	mi4 shu5	4_5	secretary	Test

c. List of surprisal sentence pairs

Item	Sentence	Pinyin	Tone	Translation	Surprisal condition	Target word position
-	天上飞着一只鹰	tian1 shang4 fei1 zhe5 yi4 zhi1 ying1	1	There is an eagle flying in the sky	low	sentence-final
-	天上肥着一只鹰	tian1 shang4 fei2 zhe5 yi4 zhi1 ying3	2	There is an eagle being fat in the sky	high	sentence-final
1	有一只鹰在天上飞	you3 yi4 zhi1 ying4 zai4 tian1 shang4 fei1	1	There is an eagle flying in the sky	low	sentence-final
1	有一只鹰在天上肥	you3 yi4 zhi1 ying4 zai4 tian1 shang4 fei2	2	There is an eagle flying in the sky	high	sentence-final
2	新买的房子还没有刷墙	xin1 mai3 de5 fang2 zi5 hai2 mei2 you3 shua1 qiang2	2	They haven't painted the walls of the new house	low	sentence-final
2	新买的房子还没有刷枪	xin1 mai3 de5 fang2 zi5 hai2 mei2 you3 shua1 qiang1	1	They haven't painted the guns of the new house	high	sentence-final
3	连日暴雨导致村庄被淹	lian2 ri4 bao4 yu3 dao3 zhi4 cun1 zhuang1 bei4 yan1	1	The village was flooded due to days of heavy rain	low	sentence-final
3	连日暴雨导致村庄被掩	lian2 ri4 bao4 yu3 dao3 zhi4 cun1 zhuang1 bei4 yan3	3	The village was covered due to days of heavy rain	high	sentence-final
4	儿童需要悉心抚养	er2 tong2 xu1 yao4 xi1 xin1 fu3 yang3	3	Children need to be nurtured carefully	low	sentence-final
4	儿童需要悉心抚秧	er2 tong2 xu1 yao4 xi1 xin1 fu3 yang1	1	Children need to be nurtured carefully	high	sentence-final
5	屠夫正在磨刀	tu2 fu1 zheng4 zai4 mo2 dao1	1	The butcher is sharpening his knife	low	sentence-final
5	屠夫正在磨稻	tu2 fu1 zheng4 zai4 mo2 dao4	4	The butcher is sharpening rice	high	sentence-final
6	农民在山坡上种了树	nong2 min2 zai4 shan1 po1 shang4 zhong4 le5 shu4	4	The farmers planted trees on the hillside	low	sentence-final
6	农民在山坡上种了书	nong2 min2 zai4 shan1 po1 shang4 zhong4 le5 shu1	1	The farmers planted books on the hillside	high	sentence-final
7	动物会随季节变化换毛	dong4 wu4 hui4 sui2 ji4 jie2 bian4 hua4 huan4 mao2	2	Animals change their fur according to different seasons	low	sentence-final
7	动物会随季节变化换猫	dong4 wu4 hui4 sui2 ji4 jie2 bian4 hua4 huan4 mao3	1	Animals change cats according to different seasons	high	sentence-final
8	他时常和朋友打赌	ta1 shi2 chang2 he2 peng2 you5 da3 du3	3	He often makes bets with his friends	low	sentence-final
8	他时常和朋友打毒	ta1 shi2 chang2 he2 peng2 you5 da3 du2	2	He often beats poison with his friends	high	sentence-final

9	他喜欢与人交谈	ta1 xi3 huan1 yu3 ren2 jiao1 tan2	2	He often beats poison with his friends	low	sentence-final
9	他喜欢予人焦炭	ta1 xi3 huan1 yu3 ren2 jiao1 tan4	4	He likes to give people charred coal	high	sentence-final
10	妈妈叫他回家吃饭	ma1 ma5 jiao4 ta1 hui2 jia1 chi1 fan4	4	Mom told him to come home for dinner	low	sentence-final
10	妈妈叫他回家吃矾	ma1 ma5 jiao4 ta1 hui2 jia1 chi1 fan2	2	Mom told him to come home for vitriol	high	sentence-final
11	他在拥挤的人群中挤掉了鞋	ta1 zai4 yong1 ji3 de5 ren2 qun2 zhong1 ji3 diao4 le5 xie2	2	Someone lost his shoe in the crowd	low	sentence-final
11	他在拥挤的人群中挤掉了血	ta1 zai4 yong1 ji3 de5 ren2 qun2 zhong1 ji3 diao4 le5 xie3	3	Someone lost his blood in the crowd	high	sentence-final
12	房间里堆满了杂物	fang2 jian1 li3 dui1 man3 le5 za2 wu4	4	The room is full of sundry goods	low	sentence-final
12	房间里堆满了杂舞	fang2 jian1 li3 dui1 man3 le5 za2 wu3	3	The room is full of sundry dances	high	sentence-final
13	锅里的包子熟了	guo1 li3 de5 bao1 zi5 shu2 le5	1	The buns in the steaming pot are ready to eat	low	sentence-medial
13	锅里的雹子熟了	guo1 li3 de5 bao2 zi5 shu2 le5	2	The hail in the steaming pot is ready to eat	high	sentence-medial
14	新建成的铁桥通车了	xin1 jian4 cheng2 de5 tie3 qiao2 tong1 che1 le6	2	The newly built iron bridge was opened for traffic	low	sentence-medial
14	新建成的铁锹通车了	xin1 jian4 cheng2 de5 tie3 qiao1 tong1 che1 le8	1	The newly built iron shovel was opened for traffic	high	sentence-medial
15	阴天出门要带雨伞	yin1 tian1 chu1 men2 yao4 dai4 yu3 san3	1	You should bring an umbrella to go outside when it is cloudy	low	sentence-medial
15	阴天杵门要带雨伞	yin1 tian1 chu3 men2 yao4 dai4 yu3 san3	3	To knock the door needs to bring an umbrella when it is	high	sentence-medial
16	记者要求发言人给出解释	ji4 zhe3 yao1 qiu2 fa1 yan2 ren2 gei3 chu1 jie3 shi4	3	The reporter requested an explanation from the speaker	low	sentence-medial
16	记者要求发言人给出街市	ji4 zhe3 yao1 qiu2 fa1 yan2 ren2 gei3 chu1 jie1 shi4	1	The reporter requested street market from the speaker	high	sentence-medial
17	小女孩向朋友挥手	xiao3 nv3 hai2 xiang4 peng2 you5 hui1 shou3	1	The little girl waved to her friend	low	sentence-medial
17	小女孩向朋友绘手	xiao3 nv3 hai2 xiang4 peng2 you5 hui4 shou3	4	The little girl drew to her friend's hand	high	sentence-medial
18	这本书印制精良	zhe4 ben3 shu1 yin4 zhi4 jing1 liang2	4	This book is welled printed	low	sentence-medial
18	这本书音质精良	zhe4 ben3 shu1 yin1 zhi4 jing1 liang2	1	This book has great sound quality	high	sentence-medial

19	她忘记水壶里还烧着水	ta1 wang4 ji4 shui3 hu2 li3 hai2 shao1 zhe5 shui3	2	She forgot the kettle was still boiling	low	sentence-medial
19	她忘记水浒里还烧着水	ta1 wang4 ji4 shui3 hu3 li3 hai2 shao1 zhe5 shui3	3	She forgot the book "the Water Margin" was still boiling	high	sentence-medial
20	戏台下的观众叫好连连	xi4 tai2 xia4 de5 guan1 zhong4 jiao4 hao3 lian2 lian2	3	There were many cheers from the audience under the	low	sentence-medial
20	戏台下的观众叫嚎连连	xi4 tai2 xia4 de5 guan1 zhong4 jiao4 hao2 lian2 lian3	2	There were howls from the audience under the stage	high	sentence-medial
21	老师教孩子们看图讲故事	lao3 shi1 jiao1 hai2 zi5 men5 kan4 tu2 jiang3 gu4 shi4	2	The teacher taught children to tell a story according to a picture	low	sentence-medial
21	老师教孩子们看兔讲故事	lao3 shi1 jiao1 hai2 zi5 men5 kan4 tu4 jiang3 gu4 shi4	4	The teacher taught children to tell a story according to a rabbit	high	sentence-medial
22	学生们喜欢做物理实验	xue2 sheng1 men5 xi3 huan1 zuo4 wu4 li3 shi2 yan4	4	The students enjoy doing physics experiments	low	sentence-medial
22	学生们喜欢做无理实验	xue2 sheng1 men5 xi3 huan1 zuo4 wu2 li3 shi2 yan5	2	The students enjoy doing nonsense experiments	high	sentence-medial
23	面条在锅里煮着	mian4 tiao2 zai4 guo1 li3 zhu3 zhe5	3	The noodles are boiling in the pot	low	sentence-medial
23	面条在锅里住着	mian4 tiao2 zai4 guo1 li3 zhu4 zhe5	4	The noodles are living in the pot	high	sentence_internal
24	病人需要及时治疗	bing4 ren2 xu1 yao4 ji2 shi2 zhi4 liao2	4	The patient needs immediate treatment	low	sentence-medial
24	病人需要及时止疗	bing4 ren2 xu1 yao4 ji2 shi2 zhi3 liao2	3	The patient needs to stop taking treatment immediately	high	sentence-medial

d. The North Wind and the Sun

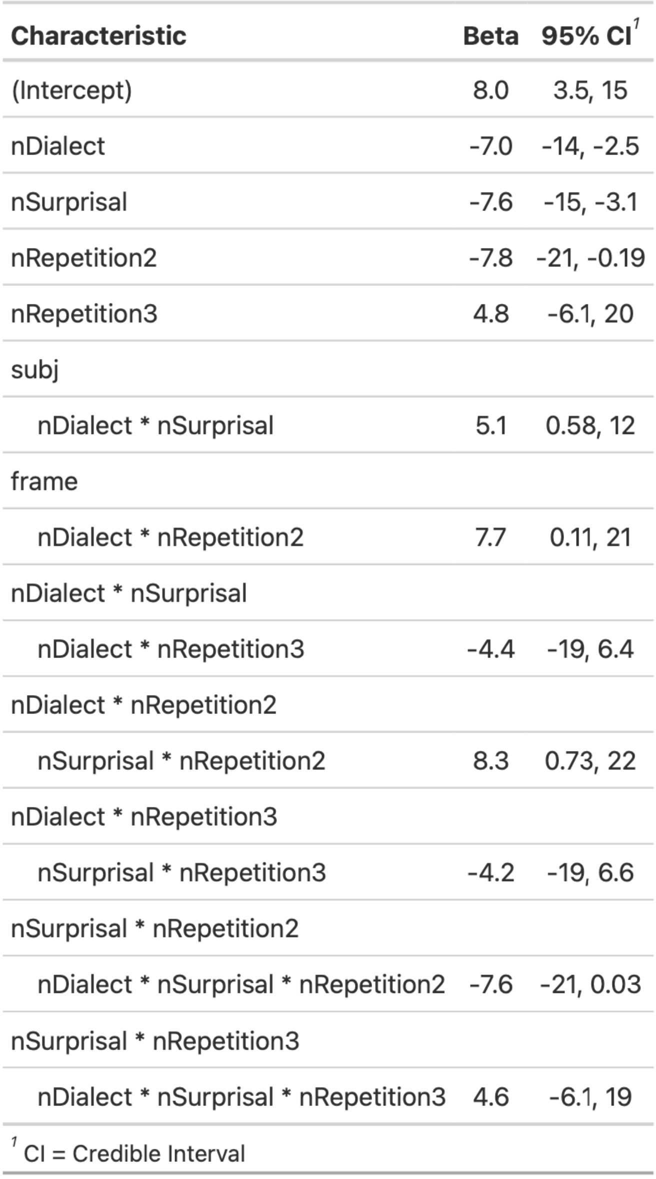
Sentences	Pinyin (in the format readable by Montreal Forced Aligner)
有一回	iou3 i4 h uei2
北风跟太阳正在那争论谁的本领大	b ei3 f e1 ng g e1 n t ai4 ia2 ng zh e4 ng z ai4 n a4 zh e1 ng l ue4 n sh uei2 d e5 b e3 n l i3 ng d a4
说着说着	sh uo1 zh e5 sh uo1 zh e5
来了一个过路的	l ai2 l e5 i2 g e4 g uo4 l u4 d e5
身上穿了一件厚袍子	sh e1 n sh a4 ng ch ua1 n l e5 i2 j ia4 n h ou4 p a2 o z ii5
他们俩就商量好了	t a1 m e5 n l ia3 j iu4 sh a1 ng l ia4 ng h ao3 l e5
说	sh uo1
谁能先叫这个过路的把他的袍子脱下来	sh uei2 n e2 ng x ia1 n j iao4 zh ei4 g e4 g uo4 l u4 d e5 b a3 t a1 d e5 p a2 o z ii5 t uo3 x ia4 l ai2
就算他的本领大	j iu4 s ua4 n t a1 d e5 b e3 n l i3 ng d a4
北风就卯足了劲儿	b ei3 f e1 ng j iu4 m ao3 z u2 l e5 j i4 n e2 r
拼命的吹	p i1 n m i4 ng d e5 ch uei1
可是	k e3 sh ii4
他吹的越厉害	t a1 ch uei1 d e5 ve4 l i4 h ai4
那个人就把他的袍子裹得越紧	n ei4 g e4 r e2 n j iu4 b a3 t a1 d e5 p a2 o z ii5 g uo3 d e2 ve4 j i3 n
到了末了	d ao4 l e5 m o4 l iao3
北风没辙了	b ei3 f e1 ng m ei2 zh e2 l e5
只好就算了	zh ii3 h ao3 j iu4 s ua4 n l e5
一会儿	i4 h uei4 e2 r
太阳出来一晒	t ai4 ia2 ng ch u1 l ai2 i2 sh ai4
那个人马上就把袍子脱了下来	n ei4 g e4 r e2 n m a3 sh a4 ng j iu4 b a3 p a2 o z ii5 t uo1 l e5 x ia4 l ai2
所以北风不得不承认	s uo3 i3 b ei3 f e1 ng b u4 d e2 b u4 ch e2 ng r e4 n
还是太阳比他的本领大	h ai2 sh ii4 t ai4 ia2 ng b i3 t a1 d e5 b e3 n l i3 ng d a4

e. Wo Chun poem

	Reading	Characters	Pinyin	Translation
line 1	original	暗梅幽闻花	an4 mei2 you1 wen2 hua1	The hidden plum smells pleasant
	homophonous	俺没有文化	an3 mei2 you3 wen2 hua4	I have had no education
line 2	original	卧枝伤恨底	wo4 zhi1 shang1 hen4 di3	Fallen willows makes me sad
	homophonous	我智商很低	wo3 zhi4 shang1 hen3 di1	I have a low IQ
line 3	original	遥闻卧似水	yao2 wen2 wo4 si4 shui3	I hear the stream in distance
	homophonous	要问我是谁	yao4 wen4 wo3 shi4 shui2	To ask who I am
line 4	original	易透达春绿	yi4 tou4 da2 chun1 lv4	The water is so clear and clean
	homophonous	一头大蠢驴	yi4 tou2 da4 chun3 lv2	a big dumb donkey
line 5	original	岸似绿	an4 si4 lv4	The reiver band is all green
	homophonous	俺是驴	an3 shi4 lv2	I am a donkey
line 6	original	岸似透绿	an4 si4 tou4 lv4	green as spring grass
	homophonous	俺是头驴	an3 shi4 tou2 lv2	I am a donkey
line 7	original	岸似透黛绿	an4 si4 tou4 dai4 lv4	green as emerald
	homophonous	俺是头呆驴	an3 shi4 tou2 dai1 lv2	I am a dumb donkey

f. The initial model results in Experiment 2 (Section 3.2)

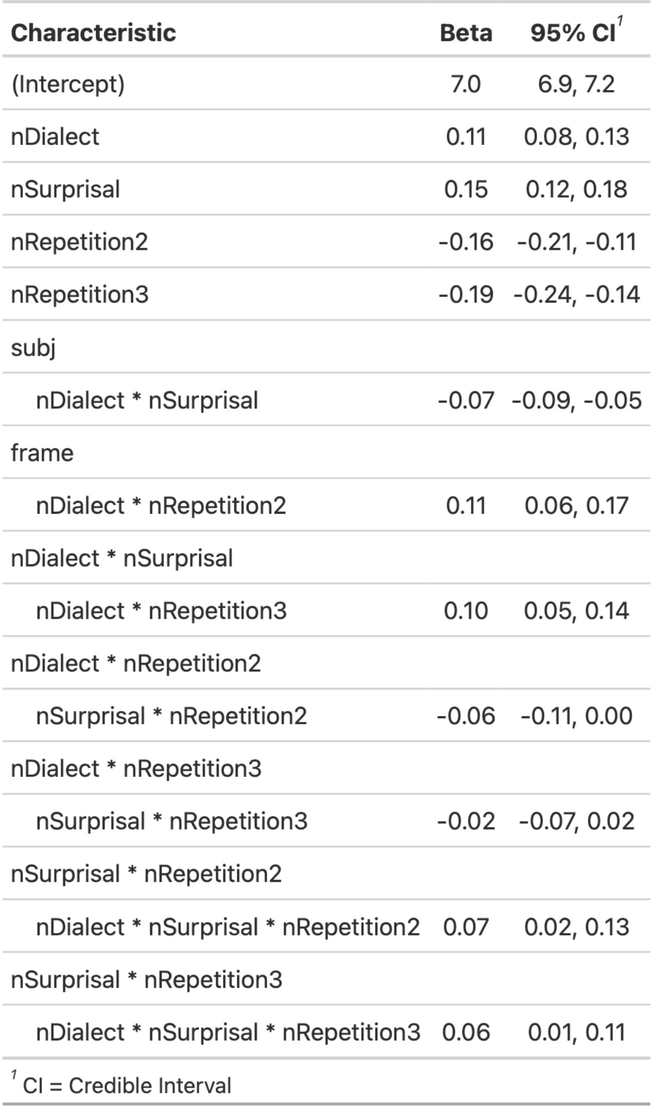
Accuracy model



Characteristic	Beta	95% CI[1]
(Intercept)	8.0	3.5, 15
nDialect	-7.0	-14, -2.5
nSurprisal	-7.6	-15, -3.1
nRepetition2	-7.8	-21, -0.19
nRepetition3	4.8	-6.1, 20
subj		
nDialect * nSurprisal	5.1	0.58, 12
frame		
nDialect * nRepetition2	7.7	0.11, 21
nDialect * nSurprisal		
nDialect * nRepetition3	-4.4	-19, 6.4
nDialect * nRepetition2		
nSurprisal * nRepetition2	8.3	0.73, 22
nDialect * nRepetition3		
nSurprisal * nRepetition3	-4.2	-19, 6.6
nSurprisal * nRepetition2		
nDialect * nSurprisal * nRepetition2	-7.6	-21, 0.03
nSurprisal * nRepetition3		
nDialect * nSurprisal * nRepetition3	4.6	-6.1, 19

[1] CI = Credible Interval

RT model

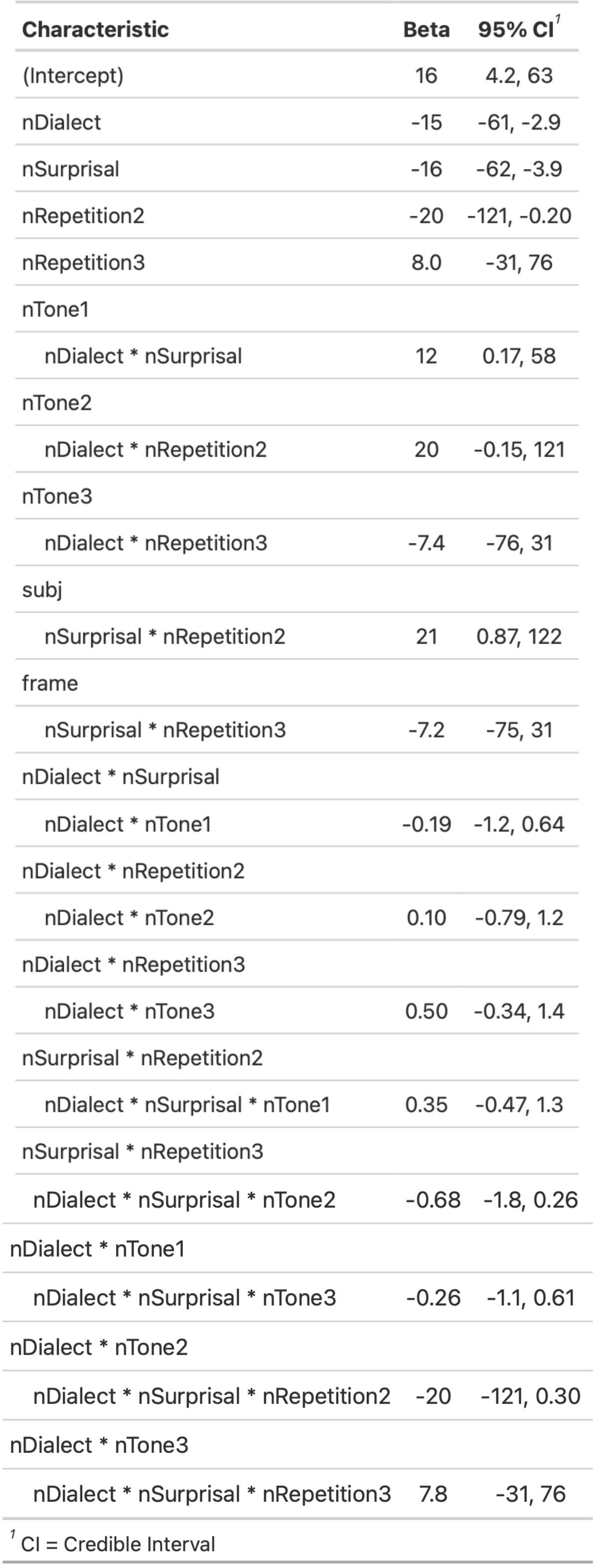


Characteristic	Beta	95% CI[1]
(Intercept)	7.0	6.9, 7.2
nDialect	0.11	0.08, 0.13
nSurprisal	0.15	0.12, 0.18
nRepetition2	-0.16	-0.21, -0.11
nRepetition3	-0.19	-0.24, -0.14
subj		
nDialect * nSurprisal	-0.07	-0.09, -0.05
frame		
nDialect * nRepetition2	0.11	0.06, 0.17
nDialect * nSurprisal		
nDialect * nRepetition3	0.10	0.05, 0.14
nDialect * nRepetition2		
nSurprisal * nRepetition2	-0.06	-0.11, 0.00
nDialect * nRepetition3		
nSurprisal * nRepetition3	-0.02	-0.07, 0.02
nSurprisal * nRepetition2		
nDialect * nSurprisal * nRepetition2	0.07	0.02, 0.13
nSurprisal * nRepetition3		
nDialect * nSurprisal * nRepetition3	0.06	0.01, 0.11

[1] CI = Credible Interval

g. The follow-up model results with the tone factors in Experiment 2 (Section 3.3)

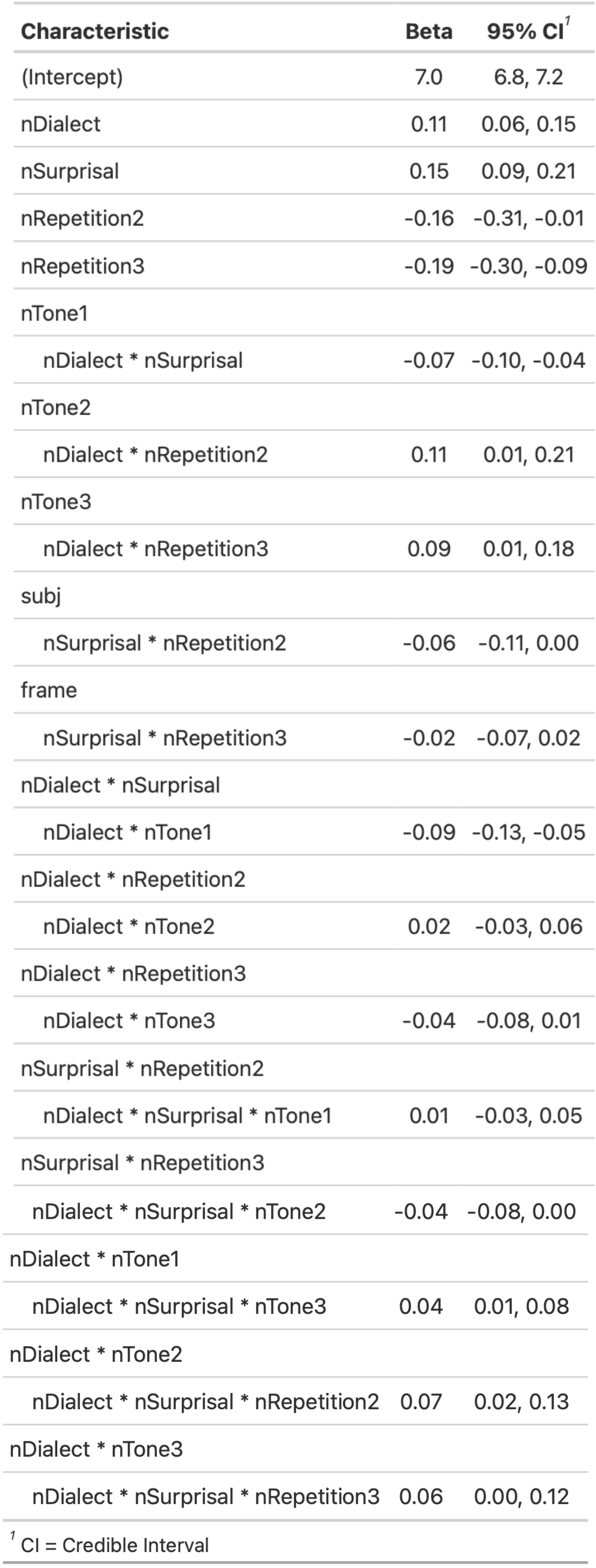
Accuracy model



Characteristic	Beta	95% CI[1]
(Intercept)	16	4.2, 63
nDialect	-15	-61, -2.9
nSurprisal	-16	-62, -3.9
nRepetition2	-20	-121, -0.20
nRepetition3	8.0	-31, 76
nTone1		
nDialect * nSurprisal	12	0.17, 58
nTone2		
nDialect * nRepetition2	20	-0.15, 121
nTone3		
nDialect * nRepetition3	-7.4	-76, 31
subj		
nSurprisal * nRepetition2	21	0.87, 122
frame		
nSurprisal * nRepetition3	-7.2	-75, 31
nDialect * nSurprisal		
nDialect * nTone1	-0.19	-1.2, 0.64
nDialect * nRepetition2		
nDialect * nTone2	0.10	-0.79, 1.2
nDialect * nRepetition3		
nDialect * nTone3	0.50	-0.34, 1.4
nSurprisal * nRepetition2		
nDialect * nSurprisal * nTone1	0.35	-0.47, 1.3
nSurprisal * nRepetition3		
nDialect * nSurprisal * nTone2	-0.68	-1.8, 0.26
nDialect * nTone1		
nDialect * nSurprisal * nTone3	-0.26	-1.1, 0.61
nDialect * nTone2		
nDialect * nSurprisal * nRepetition2	-20	-121, 0.30
nDialect * nTone3		
nDialect * nSurprisal * nRepetition3	7.8	-31, 76

[1] CI = Credible Interval

RT model



Characteristic	Beta	95% CI[1]
(Intercept)	7.0	6.8, 7.2
nDialect	0.11	0.06, 0.15
nSurprisal	0.15	0.09, 0.21
nRepetition2	-0.16	-0.31, -0.01
nRepetition3	-0.19	-0.30, -0.09
nTone1		
nDialect * nSurprisal	-0.07	-0.10, -0.04
nTone2		
nDialect * nRepetition2	0.11	0.01, 0.21
nTone3		
nDialect * nRepetition3	0.09	0.01, 0.18
subj		
nSurprisal * nRepetition2	-0.06	-0.11, 0.00
frame		
nSurprisal * nRepetition3	-0.02	-0.07, 0.02
nDialect * nSurprisal		
nDialect * nTone1	-0.09	-0.13, -0.05
nDialect * nRepetition2		
nDialect * nTone2	0.02	-0.03, 0.06
nDialect * nRepetition3		
nDialect * nTone3	-0.04	-0.08, 0.01
nSurprisal * nRepetition2		
nDialect * nSurprisal * nTone1	0.01	-0.03, 0.05
nSurprisal * nRepetition3		
nDialect * nSurprisal * nTone2	-0.04	-0.08, 0.00
nDialect * nTone1		
nDialect * nSurprisal * nTone3	0.04	0.01, 0.08
nDialect * nTone2		
nDialect * nSurprisal * nRepetition2	0.07	0.02, 0.13
nDialect * nTone3		
nDialect * nSurprisal * nRepetition3	0.06	0.00, 0.12

[1] CI = Credible Interval

REFERENCES

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, 116(5), 3099–3107.

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. (2020). Gorillas in our Midst: Gorilla. sc, a new web-based Experiment Builder. bioRxiv, 438242.

Appelbaum, I. (1996, October). The lack of invariance problem and the goal of speech perception. *Proceeding of Fourth International Conference on Spoken Language Processing*. ICSLP'96 (Vol. 3, pp. 1541–1544). IEEE.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. (1988). The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. *Perception & Psychophysics*, 44(5), 395–408.

Beddor, P. S. (2017). Speech perception in phonetics. In *Oxford Research Encyclopaedia of Linguistics*.

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological science*, 14(6), 592–597.

Besner, D., & Swan, M. (1982). Models of lexical access in visual word recognition. *The Quarterly Journal of Experimental Psychology*, 34(2), 313–325.

Besner, D., & Swan, M. (1982). Models of lexical access in visual word recognition. *The Quarterly Journal of Experimental Psychology*, 34(2), 313–325.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation hypothesis. In J. C. Goodman & H. C. Nusbaum (Eds.). *The development of speech perception* (pp. 167–224). Cambridge, MA: MIT Press.

Bird, S., Gawne, L., Gelbart, K., and McAlister, I. (2014, August). Collecting bilingual audio in remote indigenous communities. In *Proceedings of COLING 2014*, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 1015–1024).

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior research methods*, 42(3), 665–670.

Boersma, W. and Weenink, D. (2020). Praat: Doing phonetics by computer [Version 6.1.16].

Bradlow, A. R. (n.d.) ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707–729.

Bramlett, A. A., & Wiener, S. (2022). jTRACE modeling of L2 Mandarin learners' spoken word recognition at two time points in learning. *Proc. Speech Prosody 2022*, 773–776.

Bürkner, P. C. (2018). "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal*, 10(1), 395–411. doi:10.32614/RJ-2018-017.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS one*, 5(6), e10729.

Canavan, A., and Zipperlen, G. (1996). CALLFRIEND Mandarin Chinese-Mainland Dialect LDC96S55. *Philadelphia: Linguistic Data Consortium*.

Chao, Y. R. (1943). Languages and dialects in China. *The Geographical Journal*, 102(2), 63–66.

Chen, Y., & Guo, L. (2022). Zhushan Mandarin. *Journal of the International Phonetic Association*, 52(2), 309–327.

Chen, Y., Zhang, J., Chen, Y., Liu, L., Wei, J., & Dang, J. (2015, October). An articulatory analysis of apical syllables in Standard Chinese. In *2015 International Conference Oriental COCOSDA* held jointly with *2015 Conference on Asian Spoken Language Research and Evaluation* (O-COCOSDA/CASLRE) (pp. 123-127). IEEE.

Cheng, C. C. (1991). Quantifying affinity among Chinese dialects. *Journal of Chinese Linguistics monograph series*, (3), 76-110.

Chirkova, E. & Chen, Y.Y. (2011). Beijing Mandarin, the language of Beijing. HAL Id: hal-00724219.

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. Cambridge, USA. M. I. T. Press.

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(1A_Supplement), 146–146.

Chui, K., and Lai, H. L. (2008). The NCCU Corpus of Spoken Chinese: Mandarin, Hakka, and Southern Min. *Taiwan Journal of Linguistics*, 6(2).

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias?. *Perception & psychophysics*, 70, 604–618.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, 116(6), 3647–3658.

Cole, R. A. (1973). Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics*, 13(1), 153–156.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior*, 8(2), 240–247.

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory and Language*, 26(5), 527–538.

Connine, C. M., & Clifton Jr, C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human perception and performance*, 13(2), 291.

Connolly, K. (2017). Perceptual Learning. *Stanford Encyclopaedia of Philosophy*, 1.

Craik, F. I. M. & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology*, 104(3), 268–294.

Cutler, A. (2002). Lexical access. In *Encyclopaedia of cognitive science* (pp. 858–864). Nature Publishing Group.

Cutler, A., & Chen, H. C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception & Psychophysics*, 59(2), 165–179.

DawnDIY (2016). "Recorder, version 1.0.3 [smartphone application]," available at https://github.com/dawndiy/recorder (Last viewed 7/15/ 2019).

De Decker, P., and Nycz, J. (2011). For the record: Which digital media can be used for sociophonetic analysis? *University of Pennsylvania Working Papers in Linguistics*, 17(2), 51–59.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological psychology*, 1(2), 121–144.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol.*, 55, 149-179.

Duanmu, S. (1999). Metrical structure and tone: Evidence from Mandarin and Shanghai. *Journal of East Asian Linguistics*, 8(1), 1–38.

Duanmu, S. (2007). The phonology of standard Chinese. OUP Oxford.

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, 119(4), 1950–1953.

Elman, J. L., & McClelland, J. L. (1984). Speech perception as a cognitive process: The interactive activation model. *Speech and language* (Vol. 10, pp. 337–374). Elsevier.

Fernández, E. M., & Cairns, H. S. (2011). *Fundamentals of psycholinguistics*. John Wiley & Sons.

Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT journal*, 57(4), 325–334.

Flege, J. E. (1995). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, 16(4), 425–442.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing?. *Journal of Experimental Psychology: Human Perception and Performance*, 32(5), 1276.

Forster, K. I (1976). Accessing the mental lexicon. In F. Wales & E. Walker (Eds). *New Approaches to Language Mechanisms*, Amsterdam: North Holland, 257–287.

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. *The Quarterly Journal of Experimental Psychology*, 33(4), 465–495.

Forster, K. I., & Bednall, E. S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition*, 4, 53–61.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of verbal learning and verbal behavior*, 12(6), 627–635.

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.

Flynn, N. (2011). Comparing vowel formant normalisation procedures. *York Papers in Linguistics Series*, 2(11), 1–28.

Gandour, J. (1984). Tone dissimilarity judgments by Chinese listeners/声调异同辨别的测试. *Journal of Chinese Linguistics*, 235–261.

Gao, X., Yan, T. T., Tang, D. L., Huang, T., Shu, H., Nan, Y., & Zhang, Y. X. (2019). What makes lexical tone special: A Reverse Accessing Model for tonal speech perception. *Frontiers in psychology*, 10, 2830.

Gibson, J. J. (1954). A theory of pictorial perception. *Audiovisual communication review*, 2(1), 3–23.

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: differentiation or enrichment?. *Psychological Review*, 62(1), 32–41.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, 49, 585–612.

Gooskens, C., Heeringa, W., & Beijering, K. (2008). Phonetic and lexical predictors of intelligibility. *International journal of humanities and arts computing*, 2(1–2), 63–81.

Grillo, E. U., Brosious, J. N., Sorrell, S. L., & Anand, S. (2016). Influence of smartphones and software on acoustic voice measures. *International journal of telerehabilitation*, 8(2), 9.

Gui, M. C. & Liu, T. (2011). Putonghua he Beijinghua zhijian de genbenq qubie [The foundamental differences between Standard Mandarin and Beijing Mandarin]. *Yunnan Shifan Daxue Xuebao,* 9 (1), 5.

Han, H. (2000). *San Chong Men*. Tianjin Renmin Chubanshe.

Hannagan, T., Magnuson, J. S., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, 4, 563.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, 23(1), 65–94.

Hayward, K. (2014). *Experimental phonetics: An introduction*. Routledge.

He, W. (2015). Chengduhua Danzidiao de Shiyan Yuyinxue Tongji. [The experimental-phonetic analysis of Chengdu tones on monosyllables]. *Chengdu Daxue Xuebao: Shehui Kexue Ban*, 1(2015):5.

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

Ho, D. A. (2003). The characteristics of Mandarin dialects. In Thurgood & LaPolla (eds.), *The Sino-Tibetan languages*, 126–130.

Hou, J. Y. (2002). *Xiandai Hanyu Fangyan Gailun.* [The Modern Outline of Chinese Dialects]. Shanghai Education Press.

Howie, J. M. (1970). The vowels and tones of Mandarin Chinese: Acoustical measurements and experiments. [Ph.D. dissertation, Indiana University].

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones* (Vol. 18). Cambridge University Press.

Hu, H., & Zhang, Y. (2018). Path of vowel raising in Chengdu dialect of Mandarin. In *Proceedings of 29th North America Conference on Chinese Linguistics (NACCL-29)*. Columbus, Ohio, US.

Huang, X. N. & Gu, L. (2014). Chengduhua de Jichu Yuanyin Geju [The basic vowel space of Chengdu Dialect]. In *Proceedings of the 11th Phonetics Conference of China*.

Hulusic, V., Harvey, C., Debattista, K., Tsingos, N., Walker, S., Howard, D., & Chalmers, A. (2012, February). Acoustic rendering and auditory–visual cross-modal perception and interaction. In *Computer Graphics Forum* (Vol. 31, No. 1, pp. 102-131). Oxford, UK: Blackwell Publishing Ltd.

Jongman, A., Wang, Y., Moore, C. B., & Sereno, J. A. (2006). Perception and production of Mandarin Chinese tones. In P. Li, L. H. Tan, E. Bates & O. J. L. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics*, Vol. 1 (pp. 209–217). Cambridge University Press.

Jóźwik, A., & Shi, F. (2018). Analysis of monosyllabic tones in Mandarin Chinese produced by Polish students. In *Proc. 9th International Conference on Speech Prosody* (pp. 957–960).

Keating, P. A. (1996). *The phonology-phonetics interface*. UCLA Working Papers in Phonetics, 45–60.

Kinchla, R. A., & Wolfe, J. M. (1979). The order of visual processing: "Top-down," "bottom-up," or "middle-out". *Perception & psychophysics*, 25, 225–231.

Kluender, K. R., Diehl, R. L., and Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science*, 237(4819):1195–1197.

Kong, H., Wu, S., & Li, M. (2022). Hefei Mandarin. *Journal of the International Phonetic Association*, 1–22.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal?. *Cognitive psychology*, 51(2), 141–178.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, 107(1), 54–81.

Kuhl, P. K. and Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, 190(4209):69–72.

Kuhl, P. K. and Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic vot stimuli. *The Journal of the Acoustical Society of America*, 63(3):905–917.

Kuhl, P. K. and Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics*, 32(6): 542–550.

Lahiri, A., & Reetz, H. (2002). Underspecified recognition. *Laboratory phonology*, 7, 637–675.

Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, 38(1), 44–59.

Lee, W. S., & Zee, E. (2014). Chinese phonetics. In Huang, C.-T. James, Audrey Li, Y. H., and Simpson, Andrew (Ed.), *Handbook of Chinese Linguistics* (pp. 367–399). Wiley.

Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., and Messerli, J. (2020). Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. *Linguistics Vanguard*, 6(s3).

Levelt, W. J. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, 98(23), 13464–13471.

Li, Fang-Kuei. (1937). Languages and dialects. *The Chinese yearbook*. Shanghai: Commercial Press.

Li, Fang-Kuei. (1973). Languages and Dialects of China. *Journal of Chinese Linguistics*, 1, 1, 1–13.

Li, Q., Chen, Y., & Xiong, Z. (2019). Tianjin Mandarin. *Journal of the International Phonetic Association*, 49(1), 109–128.

Li, R. (2002). *Xiandai Hanyu Fangyan Dacidian.* [The Modern Dictionaries of Chinese Dialects]. Jiangsu Jiaoyu Chubanshe [Jiangsu Education Press].

Li, Y., Best, C. T., Tyler, M. D., & Burnham, D. (2020). Tone variations in regionally accented Mandarin. *Proceedings of the 21st Annual Conference of the International Speech Communication Association*, 1–5.

Li, N., & Hu, F. (2023). Vowels and Diphthongs in Chengdu Mandarin. In *Proceedings of International Congress of Phonetic Science*. Prague, Czech Republic.

Li, Y., & Wu, H. (2016). Phonemes of Xi'an Dialect and Statistic Study of Phonemic Combination Frequency. *The 4th International Conference on Management Science, Education Technology, Arts, Social Science and Economics* (pp. 613-616). Atlantis Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1), 1–36.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, 74(6), 431.

Lin, Y. H. (2002). Mid vowel assimilation across Mandarin dialects. *Journal of East Asian Linguistics*, 11(4), 303-347.

Lin, Y. H. (2014). Segmental phonology. In Huang, C.-T. James, Audrey Li, Y. H., and Simpson, Andrew (Ed.), *Handbook of Chinese Linguistics* (pp. 400–421). Wiley.

Liu, M., Chen, Y., & Schiller, N. O. (2020). Tonal mapping of Xi'an Mandarin and Standard Chinese. *The Journal of the Acoustical Society of America*, 147(4), 2803–2816.

Liu S., Samuel A. (2007). The role of Mandarin tones in lexical access under different contextual conditions. *Language and Cognitive Processes*, 22, 566–594.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the acoustical society of America*, 94(3), 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y. I., & Yamada, T. (1994). Training Japanese listeners to identify English/r/and/l/. III. Long-term retention of new phonetic categories. *The Journal of the acoustical society of America*, 96(4), 2076–2087.

Lotto, A. J., & Holt, L. L. (2016). Speech perception: The view from the auditory system. *Neurobiology of language* (pp. 185–194). Academic Press.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, 49(2B), 606–608.

Malins, J. G., & Joanisse, M. F. (2010). The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language*, 62(4), 407–420.

Marslen-Wilson, W. D. (1980, December). Speech understanding as a psychological process. In *Spoken Language Generation and Understanding: Proceedings of the NATO*

Advanced Study Institute held at Bonas, France, June 26–July 7, 1979 (pp. 39–67). Dordrecht: Springer Netherlands.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2), 71–102.

Marslen-Wilson, W. D. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148–172). The MIT Press.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1), 29–63.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, 32(3), 543–562.

McClelland, J. L., & Elman, J. L. (1986). Interactive processes in speech perception: The TRACE model. In Parallel distributed processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and biological models (pp. 58–121).

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, 88(5), 375.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception?. *Trends in cognitive sciences*, 10(8), 363–369.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746–748.

McKay, C. M. (2021). No evidence that music training benefits speech perception in hearing-impaired listeners: A systematic review. *Trends in hearing*, 25, 2331216520985678.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive science*, 30(6), 1113–1126.

McQueen, J. M., Norris, D., & Cutler, A. (2006). Are there really interactive processes in speech perception?. *Trends in Cognitive Sciences*, 10(12), 533.

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and speech*, 49(1), 101–112.

Melguy, Y. V. (2022). Perceptual learning for speech: Mechanisms of phonetic adaptation to an unfamiliar accent. [PhD dissertation, University of California, Berkeley].

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.

Mirman, D. (2017). *Growth curve analysis and visualization using R*. CRC press.

Moore, C. B., & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *The Journal of the Acoustical Society of America*, 102(3), 1864–1877.

Morton, J. (1969). Interaction of information in word recognition. *Psychological review*, 76(2), 165.

Morton, J. (1970). A functional model for memory. *Models of human memory*, 203–254.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In *Processing of visible language* (pp. 259–268). Boston, MA: Springer US.

Morton, J., Sasanuma, S., Patterson, K., & Sakuma, N. (1992). The organization of the lexicon in Japanese: Single and compound kanji. *British Journal of Psychology*, 83(4), 517–531.

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and speech*, 38(3), 289–306.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, 111(3), 721.

Nearey, T. M. (1995). A double-weak view of trading relations: comments on Kingston and Diehl. In Connell and Arvaniti (1995), 28–40.

Nearey, T. M. (1977). *Phonetic Feature Systems for Vowels*. [PhD Dissertation, University of Alberta]. Reprinted 1978 by the Indiana University Linguistics Club.

Newkline (2020). “Awesome Voice Recorder, version 1.1.2 [Android smartphone application], version 8.0.4 [iOS smartphone application],” available at http://newkline.com/ (Last viewed 1/12/2020).

Nicenboim, B., Schad, D., & Vasishth, S. (2024). *An introduction to Bayesian data analysis for cognitive science*. Under contract with Chapman and Hall/CRC statistics in the social and behavioral sciences series. https://vasishth.github.io/bayescogsci/book/

Norman, J. (2003). The Chinese dialects: phonology. *The Sino-Tibetan Languages*, 72–83.

Norman, J. (2013). *Chinese* (18th Ed.). Cambridge University Press. (1st Ed. in 1988).

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, 52(3), 189–234.

Norris, D. (1999). The merge model: Speech perception is bottom-up. *The Journal of the Acoustical Society of America*, 106(4), 2295–2295.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review*, 115(2), 357.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23(3), 299–325.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, 47(2), 204–238.

Peking University Chinese Linguistics Department. (1989). *Hanyu Fangyin Zihui*. [The Sounds of Chinese Dialects]. Wenzi Gaige Chubanshe.

Perkell, J. S., & Klatt, D. H. (Eds.). (2014). *Invariance and variability in speech processes*. Psychology Press.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *The Journal of the Acoustical Society of America*, 61(5), 1352–1361.

Pisoni, D. B. (1981). Some current theoretical issues in speech perception. *Cognition*, 10(1–3), 249.

Pisoni, D. B., & McLennan, C. T. (2016). Spoken word recognition: Historical roots, current theoretical issues, and some new directions. In *Neurobiology of language* (pp. 239–253). Academic Press.

Protopapas, A. (1999). Connectionist modeling of speech perception. *Psychological Bulletin*, 125(4), 410.

Raphael, L. J. (2021). Acoustic cues to the perception of segmental phonemes. *The handbook of speech perception*, 603–631.

Rauss, K., & Pourtois, G. (2013). What is bottom-up and what is top-down in predictive coding?. *Frontiers in psychology*, 4, 276.

Reid, A., Burnham, D., Kasisopa, B., Reilly, R., Attina, V., Rattanasone, N. X., & Best, C. T. (2015). Perceptual assimilation of lexical tone: The roles of language experience and visual information. *Attention, Perception, & Psychophysics*, 77, 571–591.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological review*, 89(1), 60.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, and Psychophysics* (Vol. 71, Issue 6, pp. 1207–1218). Springer.

Seitz, A. R., Yamagishi, N., Werner, B., Goda, N., Kawato, M., & Watanabe, T. (2005). Task-specific disruption of perceptual learning. *Proceedings of the National Academy of Sciences*, 102(41), 14895-14900.

Shi, F. (2008). *Yu Yin Ge Ju.* [Patterns in Phonetics]. Beijing: Shangwu.

Shi, F. & Wang, P. (2006). Beijing Danyinzi Shengdiao de Tongji Fenxi. [Statistical analysis of Beijing tones on monosyllables]. *Chinese Language*, 1(2006):8.

Shuai, L., & Malins, J. G. (2017). Encoding lexical tones in jTRACE: A simulation of monosyllabic spoken word recognition in Mandarin Chinese. *Behavior Research Methods*, 49, 230–241.

So, C. K., & Best, C. T. (2010). Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences. *Language and speech*, 53(2), 273–293.

Staub, A., Grant, M., Astheimer, L., and, A. C.-J. of M., & 2015, undefined. (n.d.). The influence of cloze probability and item constraint on cloze task response time. Elsevier.

Steriade, D. (1995). Underspecification and markedness. In Goldsmith, J. A. (Ed.). *Handbook of Phonological Theory* (pp. 114-174). Blackwell.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4), 1872–1891.

Stevens, K. N. (2008). Features in Speech Perception and Lexical Access. *The Handbook of Speech Perception* (pp. 124–155).

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. *Perspectives on the study of speech*, 1–38.

Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *The Journal of the Acoustical Society of America*, 55(3), 653–659.

Storms, R. L. (1998). *Auditory-visual cross-modal perception phenomena*. [Doctoral dissertation, Naval Postgraduate School].

Strauss, T. J., Harris, H. D., & Magnuson, J. S. "jTRACE: A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods*, 39(1), 19-30. 2007.

Szeto, P. Y., Ansaldo, U., & Matthews, S. (2018). Typological variation across Mandarin dialects: An areal perspective with a quantitative approach. *Linguistic Typology*, 22(2), 233–275.

Taft, M. (1986). Lexical access codes in visual and auditory word recognition. *Language and Cognitive Processes*, 1(4), 297–308.

Taft, M. (1991). Reading and the mental lexicon: Essays in cognitive psychology. L. Erlbaum.

Taft, M. (2001). Lexical access, cognitive psychology of. *International encyclopedia of the social & behavioral sciences*, 8743–8748.

Taft, M. (2006). Processing of characters by native Chinese readers. Cambridge University Press.

Taft, M. (2015). The nature of lexical representation in visual word recognition. In Pollatsek, A., & Treiman, R. (Eds.). *The Oxford handbook of reading*. Oxford University Press.

Taft, M., & Chen, H. C. (1992). Judging homophony in Chinese: The influence of tones. In *Advances in psychology* (Vol. 90, pp. 151–172). North-Holland.

Taft, M., & Hambly, G. (1986). Exploring the cohort model of spoken word recognition. *Cognition*, 22(3), 259–282.

Tang, C., & van Heuven, V. J. (2007). Mutual intelligibility and similarity of Chinese dialects: Predicting judgments from objective measures. *Linguistics in the Netherlands*. AVT Publications, 24, 223–234.

Tang, C., & van Heuven, V. J. (2008). Mutual intelligibility of Chinese dialects tested functionally. 145–156.

Tang, C., & van Heuven, V. J. (2009). Mutual intelligibility of Chinese dialects experimentally tested. *Lingua*, 119(5), 709–732.

Tang, C., & van Heuven, V. J. (2011). Tone as a Predictor of Mutual Intelligibility of Chinese Dialects. In *Proceeding of International Congress of Phonetic Science* (pp. 1962–1965).

Tang, P., Yuen, I., Xu Rattanasone, N., Gao, L., & Demuth, K. (2019). The acquisition of Mandarin tonal processes by children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, 62(5), 1309–1325.

Tupper, P., Leung, K., Wang, Y., Jongman, A., and Sereno, J. A. (2020). Characterizing the distinctive acoustic cues of Mandarin tones. *The Journal of the Acoustical Society of America*, 147(4), 2570–2580.

Vroomen, J., & Gelder, B. D. (2000). Sound enhances visual perception: cross-modal effects of auditory organization on vision. *Journal of experimental psychology: Human perception and performance*, 26(5), 1583.

Wang, X. (2013). Perception of mandarin tones: The effect of L1 background and training. *Modern Language Journal*, 97(1), 144–160.

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.

Wang, L. J., Lu, J. M., Fu, H. Q., Ma, Z. & Su, P. C. (2006). *Xian Dai Han Yu.* [Modern Chinese] (10th Ed.). The Commercial Press.

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.

Warren, R. M., & Obusek, C. J. (1971). Speech perception and phonemic restorations. *Perception & Psychophysics*, 9, 358–362.

Watanabe, T., Nanez, J. E., & Sasaki, Y. (2001). Perceptual learning without perception. *Nature*, 413(6858), 844–848.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3), 387–401.

Weil, S. (2001). Foreign accented speech: Encoding and generalization. *Journal of the Acoustical Society of America*, 109(5), 2473.

Wheeldon, L., & Waksler, R. (2004). Phonological underspecification and mapping mechanisms in the speech recognition lexicon. *Brain and language*, 90(1-3), 401-412.

Wiener, S., & Ito, K. (2016). Impoverished acoustic input triggers probability-based tone processing in mono-dialectal Mandarin listeners. *Journal of Phonetics*, 56, 38–51.

Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and speech*, 59(1), 59-82.

Wissing, D., & Pienaar, W. (2014). Evaluating vowel normalisation procedures: A case study on Southern Sotho vowels. *Southern African Linguistics and Applied Language Studies*, 32(1), 97–111.

Wu, J., Chen, Y., Van Heuven, V. J., & Schiller, N. O. (2016). Predicting tonal realizations in one Chinese dialect from another. *Speech Communication*, 76, 1–27.

Wu, J., Chen, Y., van Heuven, V. J., & Schiller, N. O. (2018). Applying functional partition in the investigation of lexical tonal-pattern categories in an under-resourced Chinese dialect. *Communications in Computer and Information Science*, 807, 24–35.

Wurm, S. A., Li, R., Baumann, T., & Lee, M. W. (1987*). Language Atlas of China*. Longman.

Wyunhe. (2011). Map of Sinitic dialect – English version. Retrieved April 30, 2022, from https://commons.wikimedia.org/wiki/File:Map_of_sinitic_languages_full- en.svg

Xia, L., & Hu, F. (2016). Vowels and Diphthongs in the Taiyuan Jin Chinese Dialect. *INTERSPEECH*, 993–997.

Xu Rattanasone, N., Tang, P., Yuen, I., Gao, L., & Demuth, K. (2018). Five-year-olds' acoustic realization of Mandarin tone sandhi and lexical tones in context are not yet fully adult-like. *Frontiers in psychology*, 9, 817.

Xu, J., & Taft, M. (2015). The effects of semantic transparency and base frequency on the recognition of English complex words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(3), 904.

Xu, Y. (1994). Production and perception of coarticulated tones. *The Journal of the Acoustical Society of America*, 95(4), 2240–2253.

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

Yang, X. H. (2011). Jinanhua de Yiji Yuanyin Geju. [The acoustic space of Jinan vowels]. *Northern Literature*.

Ye, Y., & Connine, C. M. (1999). Processing spoken Chinese: The role of tone information. *Language and cognitive processes*, 14(5-6), 609–630.

You, H., & Magnuson, J. S. (2018). TISK 1.0: An easy-to-use Python implementation of the time-invariant string kernel model of spoken word recognition. *Behavior Research Methods*, 50(3), 871–889.

Yuan, J. H. (1960). *Hanyu Fangyan Gaiyao.* [An Outline of the Chinese Dialects]. Beijing: Wenzi Gaige Chubanshe.

Yuan, J., and Liberman, M. (2014). F0 declination in English and Mandarin broadcast news speech. *Speech Communication*, 65, 67–74.

Zee, E., & Lee, W. S. (2007, August). Vowel typology in Chinese. In *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1429–1432). Saarbrücken: Saarland University.

Zhang, C. (2018). Online adjustment of phonetic expectation of lexical tones to accommodate speaker variation: a combined behavioural and ERP study. *Language, Cognition and Neuroscience*, 33(2), 175–195.

Zhang, C., Jepson, K., Lohfink, G., & Arvaniti, A. (2021). Comparing acoustic analyses of speech data collected remotely. *The Journal of the Acoustical Society of America*, 149(6), 3910–3916.

Zhang, Y., & Kirby, J. (2020). The role of F 0 and phonation cues in Cantonese low tone perception. *The Journal of the Acoustical Society of America*, 148(1), EL40–EL45.

Zhao, L. (2017). *The Neglected Tone in Chinese Word Recognition—An Account of Frequency Effect on Lexical Access*. [MA dissertation, University College London].

Zheng, Y., Sproat, R., Gu, L., Shafran, I., Zhou, H., Su, Y., ... & Yoon, S. Y. (2005, September). Accent detection and speech recognition for Shanghai-accented Mandarin. In *Interspeech* (pp. 217–220).

Zhou, H. (1999). *Vowel systems in Mandarin languages*. [Doctoral dissertation, University of Toronto].

Zhu, X., Yi, L., Zhang, T., & Nguyễn, Đ.H. (2019). Dipping tones in multi-register and four-level model. *Journal of Chinese Linguistics*, 47, 321–344.

VITA

Liang Zhao was born in October 1993 in Hebei, China. She completed a B.A. in English Language and Literature with honours from Jiangnan University in July 2016. In the following September, she joined the MA program in Linguistics at University College London under the supervision of Dr. Yi Xu and graduated with distinction in November 2017. During 2017 to 2019, she worked as a curriculum analyst for the research unit at Offcn Education Tech company in the Beijing headquarter. In September 2019, she started the PhD study at the University of York, under the guidance of Dr. Eleanor Chodroff and Dr. Paul Foulkes, and worked with Dr. Shayne Sloggett and Dr. Philip Harrison in the Department of Language and Linguistic Science. In 2024, Liang will start an Assistant Professor/Lecturer position in China.