Emerging modes of temporal coordination : Mandarin and non-native consonant clusters

Qianwen Guan

USPC
Université Sorbonne
Paris Cité

PARIS DIDEROT

Thèse de doctorat

de l'Université Sorbonne Paris Cité

Préparée à l'Université Paris Diderot

Ecole doctorale Sciences du langage 132

Laboratoire CLLILAC-ARP

Emerging modes of temporal coordination:

Mandarin and non-native consonant clusters

Par Qianwen Guan

Thèse de doctorat de Linguistique

Dirigée par Ioana Chitoran

Présentée et soutenue publiquement à Paris le 19 juillet 2019

Président du jury : M. Pierre Hallé, Directeur de recherche, CNRS

Rapporteurs : M. Alexei Kochetov, Professor, University of Toronto

Mme Marianne Pouplier, Research scientist, LMU Munich

Examinateurs : M. Tomas Lentz, Assistant Professor, University of Amsterdam

Directeur de thèse : Mme Ioana Chitoran, Professeur, Université de Paris

Titre : Emergence de nouveaux modes de coordination temporelle : le mandarin face aux clusters consonantiques

Résumé : Cette thèse examine la perception et la production des clusters consonantiques (CC) non-natifs par des locuteurs natifs du chinois mandarin, une langue à structure syllabique relativement simple. Cette étude s'est concentrée sur les différentes modifications (ou "erreurs") qui apparaissent dans la perception et la production des locuteurs, à la lumière du rôle joué par leur connaissance phonologique et par leur sensibilité aux détails phonétiques.

J'émets les hypothèses suivantes : si, tout d'abord, la connaissance phonotactique native affecte principalement l'adaptation des séquences non-natives, les locuteurs du mandarin percevront et produiront systématiquement un vocoïde dans les clusters consonantiques. En revanche, si la sensibilité aux détails phonétiques contribue principalement à l'adaptation, les locuteurs du mandarin produiront diverses modifications en fonction des propriétés phonétiques des clusters auxquels ils sont exposés. Ces hypothèses ont été testées à travers une série d'expériences : un test de discrimination ABX, un test de transcription et un test de production.

Dans le test de discrimination ABX, les locuteurs du mandarin se sont montrés très sensibles au contraste CC-CVC. Cela indique que la phonotactique native n'empêchait pas leur perception des clusters non-natifs. Les participants se sont plutôt appuyés sur les détails phonétiques des clusters. Plus le relâchement de la première consonne (dans les clusters occlusive-occlusive) était faible, moins l'épenthèse était perçue.

Dans le test de transcription, contrairement aux résultats du test de discrimination, les locuteurs du mandarin ont transcrit les CCs non-natifs avec une proportion élevée de voyelles épenthétiques. Cependant, la transcription des voyelles pourrait être influencée par l'orthographe du pinyin.

Par conséquent, nous avons mené un test de production, où les participants entendaient les stimuli contenant des clusters avant de les prononcer à voix haute. Les résultats de ce test de production ont montré que les locuteurs du mandarin produisent un vocoïde au sein des CCs, et que ce vocoïde est similaire à une voyelle réduite en mandarin, de courte durée, avec une qualité semblable à un schwa. Il est intéressant de noter que, malgré l'absence de clusters en mandarin, les locuteurs produisent parfois les clusters CC "correctement" ou avec une période de voisement, en s'appuyant uniquement sur des inputs auditifs. Mais les mesures acoustiques de ces différents types de production indiquent que le mode de coordination temporelle natif était maintenu dans la production avec vocoïdes, même si les locuteurs étaient capables de compresser le vocoïde acoustiquement. La production de clusters non-natifs par des locuteurs du mandarin est donc fortement affectée par leur connaissance phonologique, alors que leur perception de ces mêmes clusters est principalement influencée par leur sensibilité aux détails phonétiques.

Mots clefs : phonotactique, clusters consonantiques, vocoïde, perception, coordination temporelle, mandarin, non-natif

Title: Emerging modes of temporal coordination: Mandarin and non-native consonant clusters

Abstract:

This dissertation investigates the perception and production of non-native consonant clusters (CCs) by native speakers of Mandarin Chinese, a language with relatively simple syllable structure. We focus on the different modifications ('errors') that emerge in perception and production, in light of the role played by phonological knowledge and by sensitivity to phonetic details.

We hypothesized that if native phonotactic knowledge is primarily affecting non-native adaptation, Mandarin speakers will perceive and produce a vowel systematically in consonant clusters. Alternatively, if sensitivity to phonetic details primarily contributes to adaptation, Mandarin speakers will show various modifications depending on phonetic properties of the actual clusters presented. These hypotheses were tested through a series of experiments—an ABX discrimination experiment, a transcription experiment, and a prompted production experiment.

In the ABX discrimination experiment, Mandarin speakers were highly sensitive to the CC-CVC contrast, showing that native phonotactics does not impede their perception of non-native clusters. Participants relied instead on the phonetic details of the clusters. The weaker the C1 burst in stop-stop clusters, the less vowel epenthesis was perceived.

In the follow-up transcription experiment, results showed that correct transcription was absent from the data. Contrary to the discrimination results, Mandarin speakers transcribed non-native CCs with a high percentage of epenthetic vowels. However, vowel transcription may be biased by Pinyin orthography.

Therefore, we conducted a production experiment, where speakers heard the stimuli with clusters and produced them aloud. The results of this prompted production experiment showed that Mandarin speakers produce a vocoid (a 'vowel' in a purely phonetic sense, see Pike, 1943) within CCs, which is similar to a reduced vowel in Mandarin, with short duration and schwa-like quality. The acoustic measures of the production indicated that the native gestural timing pattern was maintained in the production with vocoids, even though speakers were able to compress the vocoid acoustically. Interestingly, despite the absence of clusters in Mandarin, speakers sometimes 'correctly' produced non-native CC sequences, or produced them with a period of voicing, just relying on auditory inputs. We thus learned that the production of non-native clusters by Mandarin speakers is

highly affected by their phonological knowledge, while their perception of the same clusters is primarily influenced by their sensitivity to phonetic details.

Keywords: phonotactics, consonant clusters, epenthetic vowels, perception, temporal coordination, Mandarin, non-native

Acknowledgements

I would like to extend my thanks to the people, who so generously contributed to the work presented in this dissertation.

First of all, my deep gratitude goes to my amazing advisor, Ioana Chitoran. After all these years, I still remember the day when I discussed the topic of the dissertation with her. Ioana introduced the world of phonotactics to me. She taught me how to see things differently and how to put ideas into words logically. We have spent hours and hours looking at acoustic data together. I thank Ioana wholeheartedly, not only for her marvelous academic support, but also for giving me so many wonderful opportunities. She let me attend her project meetings, showed me how to cooperate with other researchers, and encouraged me to present my work in international conferences, from Munich to Hawaii.

I am greatly honored to have Drs. Marianne Pouplier, Pierre Hallé, Alexei Kochetov and Tomas Lentz in the dissertation committee. Their work inspired me. I am deeply grateful for their valuable comments and thought-provoking questions. Special thanks go to Dr. Pouplier and Dr. Hallé, who agreed to be members of my comité de suivi. Our annual meetings have greatly contributed to pointing me in the right direction.

I would also like to express my gratitude to Dr. Fang Hu and Hongli Liang for welcoming me in their lab (the Phonetics and speech science lab at the Chinese Academy of Social Sciences), where I was able to collect the data for my dissertation.

I am grateful for the financial support I have received, without which this work would not have been possible. The doctoral fellowship from the Chinese Scholarship Council provided me with funding for the first four years of my PhD. Field data collection was supported by the research division (DRIVE) at Université

Paris Diderot. Funding from the Labex EFL " Empirical Foundations of Linguistics" allowed me to attend the LSA Summer Institute in 2015 at the University of Chicago, and thus strengthen my knowledge of different areas of linguistics.

I have been lucky enough to discuss my work with many researchers, and to learn from them. Special gratitude goes to Dr. Harim Kwon, a kind colleague and lovely friend. She gave me her valuable time whenever I approached her and helped me achieve my goal for this work. I am also grateful to Drs. Khalil Iskarous, Adamantios Gafos, Douglas H. Whalen and Jean-Yves Dommergues for their precious advice on my work and to Dr. Yair Haendler for his clear advice and help on statistical analyses.

In addition, I am grateful to Drs. Ewan Dunbar, Hiyon Yoo and Marie-Claude Paris for trusting me to TA their linguistic classes and for sharing their teaching experience.

I would like to thank my colleagues in Paris, specifically colleagues and friends in the PhD group at ARP, Hannah King, Anqi Liu, Rachel Albar, and Dr. Anisia Popescu for their incredible ideas and advice. I thank all my colleagues and friends, Darya Sandryhaila, Dr. Patricia Perez, Dr. Ismael Benali, Dr. Takeki Kamiyama, Dr. Gabriel Flambard, Prof. Georges Boulakia, and Prof. Philippe Martin for their support and encouragement.

A very special thanks goes to my friends, who are my family in France, Cécilia Xinyue Yu, Xin He, Miaoshui Wang, Valerie Van Gelder, Philippe De Boisgisson, Jacqueline Le Moal and Maxime Hirigoyen. Thank you for being always right there for me.

Finally, I would like to express my deepest gratitude to my parents and grandmother for their unconditional support and their unwavering faith in me. Even though I cannot be with you physically, and thousands of miles separate China and France, I love you deeply and think of you all the time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Cross-linguistically, languages differ with respect to the combinations of sounds they permit within and across syllables. For example, both phonemes /k/ and /t/ exist in Mandarin and Russian, but their combinations, /kt/ or /tk/, are only allowed in Russian (e.g., [kto] 'who'; ['votkə] 'vodka'), not in Mandarin.

While such phonotactic restrictions are language-specific, typologically preferred patterns can be identified. Maddieson (2013) categorizes languages into three types according to their syllable structure: simple, moderately complex, and complex. The geographical distribution of the three types of syllable structure is illustrated in Figure 1.1, based on data from the World Atlas of Language Structures (WALS) by Dryer and Haspelmath (2013, available online at http://wals.info/), which itself builds on Maddieson's UCLA Phonological Segment Inventory Database (UPSID, 1984, available online at http://web.phonetik.uni-frankfurt.de/upsid.html). The WALS contains 484 languages, for which both consonant inventory size and syllable structure are recorded.

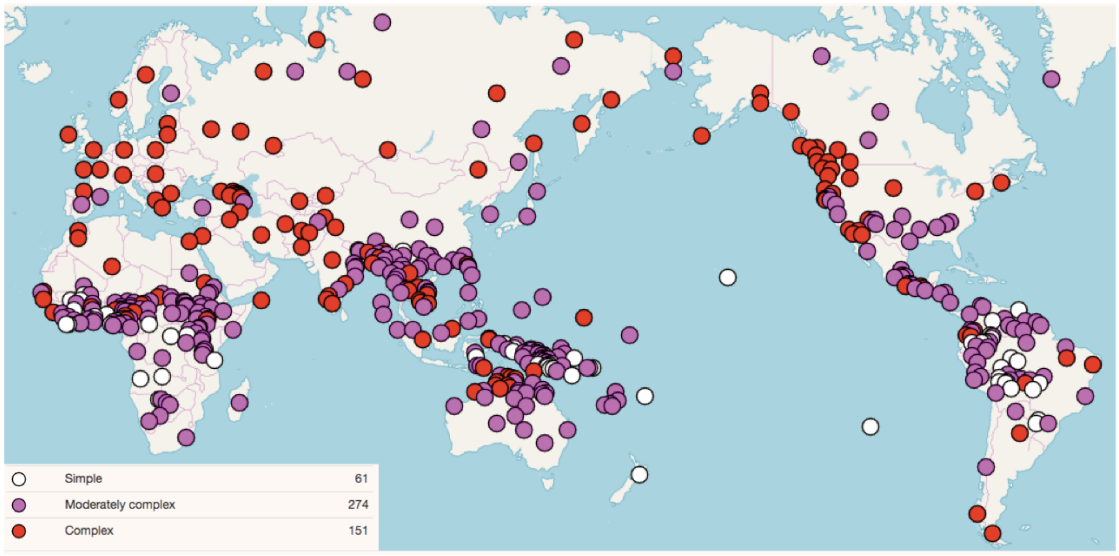


Figure 1.1: Geographical distribution of the three types of syllable structure in the WALS database across different languages: simple (white dots), moderately complex (pink dots), complex (red dots)

According to Maddieson's classification, languages with **simple** syllable structure are those containing only CV syllables, which consist of one consonant (C) followed by one vowel (V). Such syllables occur in every language, and in a small number of languages, they are the only possible type of syllable structure. This is the case, for example, in Austronesian languages like Hawaiian or Fijian, or in the Niger-Congo language Mba. Such languages make up only 12.5% of the WALS language database.

A language that allows a consonant to be added to the initial or final position of the syllable (CCV, CVC), where the second consonant is a liquid or a glide, is considered to have a **moderately complex** syllable structure. Such languages include Mandarin Chinese, Japanese or Darai (an Indo-Aryan language spoken in Nepal). The moderately complex structure is the most common structure type in the database, making up well over half of the sample (56.5%).

Languages that allow consonant sequences with more than three consonants in word-initial position, or two or more consonants in word-final position (e.g., CCCVCC), are considered to have **complex** syllable structure. Languages with complex syllable structure include English, French and German, among others, and make up 31% of the sample.

More recently, Easterday (2017) proposed an additional syllable structure category, considering that some languages have **highly complex** syllable structure. Specifically, Easterday proposes that the 'complex syllable structure' category may be further divided into those with complex syllable structure on the one hand, and highly complex syllable structure on the other. Languages with highly complex syllable structure allow more than three obstruents, or more than four consonants of any kind in word-initial or -final sequences; and/or onset or coda sequences with more than three syllabic obstruents. An example of a language with highly complex syllable structure is Itelmen, for instance /qsaɬtxt͡ʃ/ 'follow'. According to Easterday (2017), between 5% and 10% of the world's languages have highly complex syllable structure.

Examples of languages classified according to the four types of syllable structure distinguished above are shown in Table 1.1.

Table 1.1: Four categories of syllable structure. The Itelmen example is from Easterday (2017)

Structure	Simple	Moderately complex	Complex	Highly complex
Languages	Hawaiian	Mandarin Chinese	English	Itelmen
IPA	/ka:/	/kʷaŋ/	/skɹɪpt/	/qsaɬtxt͡ʃ
Translation	'to hit'	'light'	'script'	'follow'
Syllables	CV:	CGVN	CCCVCC	CCVCCCC

The central question of interest to this dissertation is how native speakers of languages with simple or moderately complex syllable structure (in our case,

Mandarin, a language with moderately complex syllable structure according to the classification in WALS) perceive and produce non-native consonant clusters (CC), a more complex type of syllable structure that does not exist in their native language. Earlier studies addressing this question found that non-native CCs were always modified to form legal sequences in the native language. One possible modification involves inserting a vowel between the two consonants (Dupoux et al., 1999; Berent et al., 2008; Hall, 2006; Broselow, 2015; de Jong and Park, 2012, among others). Such vowels inside non-native CCs have been investigated in both perception and production, but it remains unclear whether they are produced systematically. Can speakers ever produce non-native CCs 'correctly' without inserting any vowels? How do the flanking consonants coordinate with one another in the presence or absence of such acoustic vowels?

This study examines the perception and production of non-native CCs in the presence and absence of 'inserted vowels'. In perception, following Dupoux et al. (1999) such vowels are considered epenthetic vowels, and are treated as phonological segments inserted as adaptations of syllable structures that are unattested in a speaker's native language. In production, these acoustic vowels have received two different treatments. Under one view they are considered epenthetic vowels (Broselow, 1983; Hancin-Bhatt and Bhatt, 1997; Davidson et al., 2004); under another view, they are considered acoustic transitions between consonants (Davidson, 2005, 2006a). It is difficult to tell whether such acoustic vowels in production are actually the result of phonological modification, or just a by-product of an open transition between two consonants. The two types of vowels may even be mixed in the production of non-native CCs (Shaw and Davidson, 2011). We therefore need a neutral term to refer to such vowels. We will refer to them as '**epenthetic vowels**' in the perception of non-native CCs and '**vocoids**', a term in the sense of a phonetic vowel (see Pike, 1943), in the production.

In the next chapters, we aim to, first, gain a deeper understanding of the perception and production of non-native consonant clusters and, secondly, to explore how consonant gestures may coordinate in the presence and absence of vocoids between two consonants:

- In the first chapter, we review earlier studies on the perception and production of non-native CCs. Three important aspects of the perception of non-native CCs will be discussed: the role of phonotactic knowledge, the role of universal factors (sonority), and the role of phonetic factors. Regarding production, we will focus on recent studies which investigate the status of vocoids inside non-native CCs and the temporal coordination of the flanking consonants with the vocoids (Zsiga, 2003; Yanagawa, 2006; Davidson, 2006a). We will then give an overview of the phonology of the language under study, Mandarin Chinese, before introducing the hypotheses put forward in this dissertation.

- In the second chapter, we will investigate the perception of non-native CCs by Mandarin speakers using an ABX discrimination experiment, in order to understand whether Mandarin speakers are sensitive to the presence or absence of a vowel between two consonants.

- In the third chapter, we will discuss the results of an experiment in which Mandarin speakers were asked to transcribe non-native CCs after hearing them. This experiment aims to determine how often Mandarin speakers perceive a vowel between two consonants and if any additional modifications occur to repair illegal phonotactics, for example deletion of a consonant or feature changes.

- In the fourth chapter, we report on a production experiment that tested how often and how systematically Mandarin speakers produce such vocoids in

non-native CCs. Do they produce them even in a voiceless context? Can they produce CCs 'correctly' without vocoids? If so, how do the two consonant gestures coordinate in the presence vs. absence of such a vocoid?

- In the last chapter, we sum up the key findings in relation to the research questions addressed throughout this dissertation.

1.2 Studies on the perception of non-native CCs

Native speakers usually have difficulties with consonant clusters that never occur in their native language. They may misperceive non-native clusters and, when pronouncing them, often modify them by inserting a vowel between two consonants, deleting one of the consonants, or changing consonant features. Such misperceptions and/or modifications, which will be presented in section 1.2, were observed with speakers of Japanese, Spanish, English, Mandarin, among others (Polivanov, 1931; Dupoux et al., 1999; Hallé et al., 2014; Lentz and Kager, 2015; Durvasula and Kahng, 2015). These studies attribute speakers' perception difficulties to their phonotactic knowledge. For other researchers, however, how often listeners perceive a vowel inside of non-native CCs is determined by universal language restrictions on sonority (Berent et al., 2007; Berent et al., 2009). Cross-linguistically, preferred syllable onsets have rising sonority, and preferred syllable codas have falling sonority (Greenberg, 1965). Berent and colleagues found that speakers of different languages—even those without complex onsets—are also sensitive to this preferential hierarchy of onset clusters. Misperceptions have also been attributed to differences in the phonetic details of the input consonant clusters (Davidson and Shaw, 2012; Wilson et al., 2014). Wilson et al. (2014) found that native speakers of English are sensitive to the duration and intensity of the release burst of the first consonant in C1C2 consonant clusters. Specifically, a longer duration or higher amplitude of C1 release burst are both predictors of perceptual epenthesis. This led Wilson et al. to suggest that listeners perceive the release and aspiration of voiceless consonants as consonants followed by devoiced vowels. Zhao and Berent (2016) propose that the same vowel devoicing effect is involved in the perception of non-native CCs by native speakers of Mandarin.

In the following sections, we will review previous studies of the three main

factors that have been argued to affect the perception of non-native consonant clusters: language-specific phonotactic knowledge (section 1.2.1), universal language knowledge (section 1.2.2) and phonetic factors (section 1.2.3).

1.2.1 The role of language-specific phonotactic knowledge

It is well established that perceptual illusions often occur when listeners are exposed to sound sequences that are not allowed in their native language. Dupoux et al. (1999) found that Japanese listeners tend to perceive non-native consonant sequences with an illusory vowel between consonants, e.g., /abge/ is likely to be heard as /abɯge/. The vowel /ɯ/, a close back unrounded vowel, serves to modify the illicit consonant sequence /bg/ in Japanese. In an identification experiment, Dupoux et al. observed that Japanese listeners perceived a vowel between the consonants of such sequences more than 60% of the time, whereas French listeners (forming the control group) did not perceive a vowel in the same stimuli. Davidson (2007) observed the same phenomenon with American English listeners, who perceived a schwa-like vowel /ə/ inside Russian consonant clusters that are illicit in American English. Spanish provides yet another example of vowel insertion as a perceptual strategy. For instance, Hallé et al. (2014) found that native speakers of Spanish perceived the non-native sequence /spid/ as /espid/. In this case, instead of a vowel being inserted between two consonants, the vowel /e/ was perceived before a cluster beginning with /s/. Even though such clusters do exist in Spanish, they are illicit as syllable onsets, leading Spanish speakers to hear a prosthetic /e/. Hallé et al. then compared the results of Spanish listeners to those of French listeners, and found that French listeners perceived word-initial /sC/ clusters correctly, which is consistent with the fact that they are allowed as onset clusters in French. The phenomenon of prosthetic /e/ modification observed with Spanish speakers suggests that language-specific phonotactic knowledge plays a

role in perception.

However, instead of inserting a vowel between the two consonants of a cluster, listeners may also change or delete one of the consonants (Pitt, 1998; Hallé et al., 1998; Moreton, 2002). Pitt (1998) found that English listeners tended to perceive 'r' after a stop consonant /t/ when hearing a sound intermediate between /r/ and /l/ in the stimulus /t?i/. But when the stimulus was /s?i/, they perceived the sound /l/. Similarly, (Hallé et al., 1998) investigated how French listeners perceived and transcribed non-words beginning with the clusters /tl/ and /dl/ (e.g., 'tlabdo'), both of which are illicit at syllable onset in French. Results showed that French listeners sometimes misperceived the non-native tokens with word-initial /tl/ as beginning with the native sequence /kl/ instead, and likewise word-initial /dl/ tended to be perceived as native /gl/. However, the results might have been influenced by an orthographic bias: French speakers may have transcribed the illicit sequences as /kl/ and /gl/ because the corresponding combinations of letters, namely "cl" and "gl", are valid sequences at word onset in French orthography, whereas "tl" and "dl" are not.

To avoid this bias, Hallé et al. (1998) used a forced-choice identification task. Participants were instructed to identify the cluster-initial consonant by circling one of three or four letters proposed for the stimulus (P, T, K; B, D, G or D, T, G, K) after listening to it. Even so, there was a fair amount of confusion regarding the place of articulation of the first consonant: Participants misperceived the initial dental stop in /tl/ as /k/ 63.5% of the time, and the initial dental stop in /dl/ as /g/ 47.9% of the time.

In addition to various modification strategies, the quality of this inserted vowel can be influenced by the phonological knowledge of one's native language. Durvasula et al. (2018) tested specifically the quality of the inserted vowels in Mandarin Chinese and found that more than one type of vowels can be inserted by native

speakers of Mandarin between two consonants to modify illicit syllable structures. Durvasula and colleagues argued that the choice of vowels is a two-step process: first, phonological alternations (phonetic knowledge about the surface representations) are matched to acoustic details in the inputs; then, the listeners' decision on the quality of vowels is biased towards a vowel that is permitted by phonotactic restrictions in their native language (for instance, C is only allowed next to V). According to the authors, the best vowel to match phonetic properties inside consonant clusters should be a reduced or deleted vowel in the surface representations (/V/→ Ø). As for Mandarin, the best candidate should be a central vowel /ə/, which occurs in toneless or unstressed syllables in Mandarin. However, the reduced vowel /ə/ does not occur as an epenthetic V in all contexts, as some consonants are not allowed next to /ə/. For instance, in Mandarin, the alveopalatal consonant /tɕʰ/ can be followed by the vowel /i/ (/tɕʰi/), but not the schwa /ə/ (*/tɕʰə/). Therefore, Mandarin listeners should be more likely to perceive /i/ in illicit clusters beginning with /tɕʰ/. In such cases, phonotactic restrictions influence the quality of vowels.

To verify these predictions, Durvasula et al. (2018) used two experiments in their study: first, an ABX discrimination experiment, in which participants heard three stimuli A, B and X in succession, and then had to report whether X was similar to A or to B. Second, an identification experiment, in which participants were asked to determine whether the stimuli contained /i/, an other vowel, or no vowel at all. The results, consistently with their predictions, showed that participants perceived a vowel between two consonants, but the quality of the vowel varied across different phonotactic contexts. A schwa /ə/ was perceived more often after /tʰ/ in a non-word with consonant clusters, e.g., /atʰma/ → /atʰəma/, however, a vowel /i/ was perceived following an alveopalatal consonant /tɕʰ/, e.g., /atɕʰma/ → /atɕʰima/. For the control group of native speakers of English, the vowel was always

identical, a result already observed by Davidson (2007), as mentioned above. This shows that the Mandarin speakers' performance was due to their phonological alternations (surface representations) and phonotactic knowledge of their native language.

An interesting observation from the results of Durvasula and Kahng's identification experiment is that 60% of the time at least, Mandarin participants reported they did not perceive a vowel. This is a much higher percentage than that of Japanese listeners in the identification task used by Dupoux et al. (1999), who reported not hearing a vowel only 35% of the time. Moreover, in an ABX experiment carried out by Durvasula et al., Mandarin listeners were able to perceive the difference between the test tokens and their controls (e.g./at^{h}ma/ vs. /at^{h}əma/) at least 60% of the time.

Do the 'correct' responses indicate that Mandarin participants can accurately perceive the non-native consonant cluster /t^{h}m/ in /at^{h}ma/? Or did other factors help them perceive the contrast between /at^{h}ma/ and /at^{h}əma/? Durvasula et al. argue that inserted vowels may not be the only way to modify illicit consonant clusters. Other modifications, such as consonant deletion (e.g./at^{h}ma/ → /a_ma/), feature change (e.g., /at^{h}ma/ → /ap^{h}ma/), among others, may have occurred. Moreover, additional factors we have not yet systematically investigated, such as other phonetic details, may also influence the perception of non-native CCs. It is also worth noting that the Mandarin speakers tested in Durvasula et al. (2018) were learners of English, who had spent an average of 2.4 years in the US, and could be assumed to have sufficient familiarity with English clusters to know that /tm/ is possible in word-medial position in English (e.g., /ˈætməsfɪr/ 'atmosphere').

To summarize, Durvasula et al. (2018)'s study brings evidence in support of the role of language-specific phonological knowledge in the occurrence of vowels. The quality of the vowel in their participants' production was modulated by the phonol-

ogy of their native language, as it was influenced both by phonological alternations (reduced/deleted vowels) and phonotactic restrictions (phonemic combinations). However, additional factors may have influenced the percentage of responses reporting a vowel. In the next subsections, we examine the evidence for the role of universal language knowledge and phonetic factors.

1.2.2 The role of universal language knowledge

The role of universal sonority restrictions in the perception of non-native onset consonant sequences was investigated by Berent and colleagues (Berent et al., 2007, 2008, 2009, 2012a,b). In a sonority scale with fixed universal values, as proposed by Selkirk (1984) and Clements (1990), non-syllabic segments consist of the four major classes: obstruents (O), nasals (N), liquids (L) and glides (G), ranked from the least sonorous to the most sonorous: O < N < L < G. Cross-linguistically, it has been observed (Greenberg, 1965) that preferred syllable onsets have rising sonority, whereas preferred syllable codas have falling sonority. In addition, preferred onsets have a larger sonority rise, meaning a greater distance in sonority between the outer edge of the syllable and the nucleus. As a result, the following preferential hierarchy has been established by Smolensky (2006): large rises > small rises > plateaus > falls, ('>' denotes 'preferred over'). Preferred onset consonant clusters according to this scale are as follows: bl > bn > bd > lb.

The studies by Berent and colleagues (Berent et al., 2007, 2008, 2009, 2012a,b) have shown that speakers of different languages are sensitive to this preferential hierarchy of onset clusters, even for clusters that are unattested in their native language. Regardless of their familiarity with clusters, listeners tend to repair and misidentify clusters with a small sonority distance, especially with sonority reversals. For instance, the dispreferred onset sequence /lb/ with falling sonority is most likely to be confused with /ləb/, significantly more so than the high

rising sonority sequence /bl/. This has been observed with speakers of various languages, among which English (Berent et al., 2007), Korean (Berent et al., 2008) and Spanish (Berent et al., 2012a,b). The authors of these studies concluded that their findings can be attributed to universal language knowledge.

However, in a review of the study by Berent et al. (2007), Peperkamp (2007) argued that epenthesis is just one of several possible strategies for modifying non-native consonant clusters. It is not the only possible form of perceptual modification. She points out that while all the experiments in Berent et al.'s study specifically tested whether listeners perceived a vowel in different types of non-native clusters, listeners might also have changed or deleted one of the consonants, or inserted a vowel before the consonant sequences, if they had been given those options. The role of phonetic properties of the stimuli must therefore be examined very closely.

1.2.3 Phonetic factors

Acoustic and phonetic details of the inputs play an important role in the perception and production of non-native sounds. The traditional second language models—the Perceptual Assimilation Model (Best, 1995) and the Speech Learning Model (Flege, 1995)—focus on the role of 'phonetic decoding' by second language learners (L2). These models agree that L2 learners tend to match non-native sounds to the most phonetically similar native sound structures. Peperkamp and Dupoux (2003) and Peperkamp (2007) emphasize the importance of phonetic decoding in the cognitive process of perceiving non-native inputs mentioned earlier. Peperkamp and Dupoux (2003) hypothesize that non-native sounds, or inputs, should be mapped onto the most similar phonetic category, or 'phonetic surface form', in the native language. The authors argued that the choice between inserted vowels and consonant deletion, or other phonological equivalent modifications, is determined by

'phonetic minimality'. For instance, in Cantonese loanwords from English, both epenthesis and deletion may be used to repair illicit consonant clusters. However, different illicit coda consonants receive different modifications. Fricative consonants in codas lead to more epenthesis, whereas stop consonants in codas lead to deletion. This is because English fricatives have stronger phonetic cues than stops, which are often unreleased (Silverman, 1992; Yip, 1993), which is close to an empty coda. In addition to various modification strategies, the choice of the inserted vowel may also be influenced by 'phonetic minimality'. In loanword adaptation, the vowels inserted between two consonants are those that are closest to Ø, where Ø indicates 'no vowel'. For instance, the most common inserted vowel in Japanese is a close back unrounded vowel /ɯ/, which is the shortest vowel in the Japanese vowel inventory, and can be devoiced in certain contexts.

Responding to Peperkamp and Dupoux's call, Wilson et al. (2014) investigated systematically listeners' sensitivity to the phonetic properties of the stimuli in the adaptation of non-native consonant clusters. The authors tested how the same English speakers perceive and produce non-native consonant clusters. The participants heard and repeated target stimuli produced by a Russian-English bilingual speaker. Wilson et al. (2014) showed that phonetic cues of the stimuli may also affect the presence and frequency of inserted vowels. The authors tested how native speakers of English perceived and produced non-native consonant clusters, using target stimuli produced by a Russian–English bilingual speaker. Three crucial phonetic cues of the target stimuli were manipulated: the burst duration of the first consonant (C1) in C1C2 consonant clusters, the burst amplitude of C1, and pre-obstruent voicing. The task was a repetition task, in which participants were asked to repeat what they heard.

The authors observed that a longer burst duration of C1 made English speakers more likely to insert a vowel between the two consonants. When the burst duration

of C1 was 20ms long, participants repeated the consonant clusters that they heard with a vowel 20% of the time. As the burst duration increased to 50ms, they repeated consonant clusters with an added vowel 40% of the time. Wilson et al. suggest that participants may interpret the stimuli with long burst followed by aspiration as a devoiced vowel, since devoiced vowels have similar acoustic features.

In addition to burst duration, the burst amplitude of C1 can also affect the rate of vowel insertion. Wilson et al. observed that a higher amplitude of C1 resulted in more inserted vowels. In contrast, a lower amplitude of C1 reduced the rate of vowel insertion. In that case, speakers often changed the place of articulation of C1 (/tm/ →/km/) or deleted C1 (/tm/ → /_m/).

Moreover, vowel insertion occurred more often in the context of voiced onset consonants than in voiceless contexts. Wilson et al. attribute this finding to the fact that the voiced bursts have a periodic waveform, which is acoustically similar to a vowel. The longer the duration of periodicity, the more likely listeners are to perceive the voiced burst as an inserted vowel.

The perception of a long burst as a devoiced/inaudible vowel was also proposed as a possible repair strategy by Zhao and Berent (2016) in a study of Mandarin. They observed that Mandarin listeners were highly sensitive to the burst duration and intensity of C1 stops. Results showed that participants perceived a vowel between the consonants of the cluster /pl/, a universally preferred sequence, at the same rate as in /lp/, which is universally dispreferred because of the sonority reversal. Mandarin has moderately complex syllable structure and lacks consonantal onset clusters, except for consonant + glide combinations, and is also known to have non-low vowel devoicing after an aspirated consonant in the low tone context (Duanmu, 2007). Zhao and Berent (2016) suggest that Mandarin listeners may perceive an onset stop consonant with a longer burst duration as a stop consonant followed by a devoiced vowel, since the long burst shows the aperiodic energy

characteristic of aspiration. Zhao and Berent concede that, for Mandarin speakers, the sonority effect could be masked by the listeners' heightened sensitivity to phonetic properties. This suggests that, at least for Mandarin native speakers, phonetic knowledge of the native language may play a greater role than universal language knowledge in the perception of non-native consonant clusters. However, Zhao and Berent (2016) did not systematically test how phonetic knowledge influences the perception of non-native consonant clusters, since, as suggested in Wilson et al. (2014), various modifications other than insertion may occur due to phonetic properties of the input. In the current study, we tested systematically the effect of phonetic properties on the perception of non-native consonant clusters by monolingual speakers of Mandarin. We focused on the burst release of C1, including duration and intensity.

To sum up, language-specific phonotactic knowledge plays an important role in the adaptation of non-native consonant clusters. Listeners often insert a vowel between two consonants, delete one of the consonants or change its features when a particular consonant cluster is phonotactically illicit in their native language. However, the effect of phonotactic knowledge may be modulated by universal sonority restrictions, as well as by phonetic factors, such as burst duration of C1 and intensity of C1. We saw that the longer the burst duration, the higher the likelihood that an inserted vowel would be perceived and produced. But shorter and lower amplitude of C1 may result in consonant deletion and change. In the case of Mandarin specifically, we saw that listeners may perceive a reduced vowel inside illicit consonant clusters, or even interpret a long burst followed by aspiration as containing a devoiced vowel. Indeed, the inputs with nothing (shown as Ø) between two consonants are mapped onto the most similar phonetic categories, which will be a reduced/devoiced vowel (V) between two consonants(CØC → CVC). The study we propose takes into account the effect of both **language-**

specific phonological knowledge and phonetic factors (release duration and intensity of C1) in the perception and production of non-native consonant clusters. In the next section, we will review previous studies on temporal coordination in the production of non-native consonant clusters, analyzed within the framework of Articulatory Phonology.

1.3 Studies on production of non-native CCs

It is well established that coarticulation patterns are language-specific. Successive sounds, like consonant-vowel and consonant-consonant, have their own coarticulation patterns in the production of native languages (Öhman, 1966; Browman and Goldstein, 1990). But less is known about the coarticulation patterns (i.e. temporal coordination patterns) of non-native consonant clusters. What we have learned from the few previous studies is that speakers may have difficulty reproducing correctly the coarticulation pattern of the target language (Davidson, 2005, 2006a). However, the more familiar they are with non-native consonant clusters, the more closely their coarticulation patterns will resemble those of the target non-native consonant clusters (Yanagawa, 2006; Luo, 2017).

Alternatively, inaccurate coarticulation patterns in non-native consonant clusters may result in an vocoid present between two consonants. This vocoid is an open 'transition', with no vocalic segment added, which is a by-product of the inaccurate temporal coordination. Such a vocoid is different from the vocoid we presented earlier, reported in perception studies, which is a phonological segment perceptually inserted as a repair of illicit consonant clusters. Therefore, we should make it clear that in the production of non-native consonant clusters, the vocoids between two consonants may be either a lexical vowel, or an open transition. We will review coarticulation patterns of non-native consonant clusters in detail in the

current subsection.

We first introduce the notion of articulatory gesture and of temporal coordination between gestures. We do so using the framework of Articulatory Phonology (Browman and Goldstein, 1986, 1992a), because the concepts used here have been developed in detail within this model. Then, we review previous studies on temporal coordination of non-native consonant clusters.

1.3.1 Temporal coordination

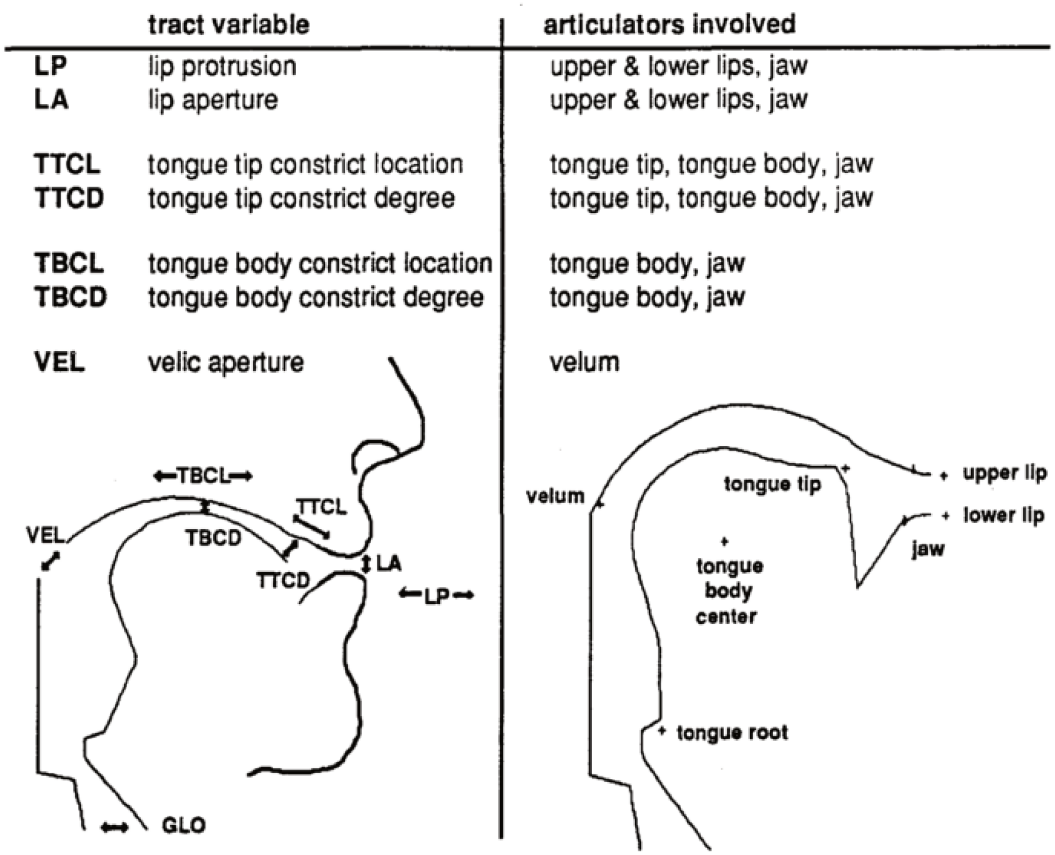


Figure 1.2: Tract variables and associated articulators (from Browman and Goldstein, 1992b)

In Articulatory Phonology, the basic units of speech are gestures, articulatory actions of different organs in the vocal tract which function as units of contrast among lexical items (Browman and Goldstein, 1986, 1992a). Gestures are dynamically specified in terms of space and time, and are characterized by three parameters.

The first is the nature of the articulators involved, including lips, tongue tip, tongue body, velum, and glottis. The second is the spatial information about the target constrictions of the articulators. Figure 1.2 from Browman and Goldstein (1992b) shows the articulators on the right with their target constrictions on the left. Each constriction is characterized by two variables specifying its spatial goals: constriction location (CL) and constriction degree (CD). The CL variable takes into account the place of the constriction in the vocal tract (labial, dental, alveolar, postalveolar, palatal, velar, uvular, and pharyngeal). The CD variable ranges from 'closed' to 'wide' with five values: closed, critical, narrow, mid, and wide. For instance, for the consonant /p/, the CL is *labial*, and the CD is *narrow*.

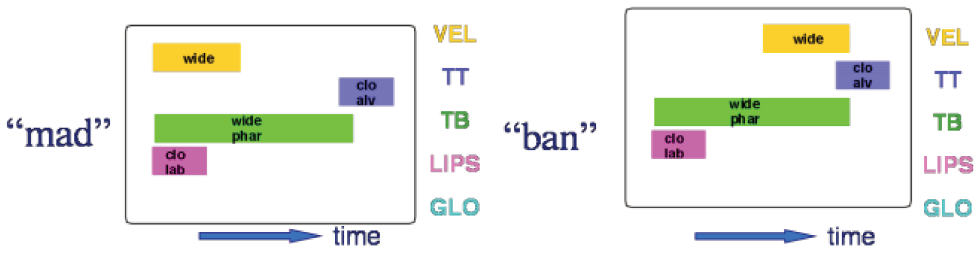


Figure 1.3: Gestural Scores for 'mad' and 'ban' (from Goldstein et al., 2009)

In the speech production system, gestures are subject to overlap. Gestural overlap is presented in a gestural score, as in Figure 1.3. The gestural scores of the two English words 'mad' and 'ban' in Figure 1.3 show that the temporal organization of the velum lowering gesture captures the difference between the two lexical items. The velum lowering gesture (in yellow) involves air escaping through

the nose to produce a nasal consonant. For the word 'mad', this gesture starts earlier than for 'ban'. Thus, gestures are 'glued' together to ensure the temporal stability of the information pattern. The information pattern in speech production plays an important role in perception. For example, the velum lowering gesture achieving its constriction goal early ensures that 'mad' is not perceived instead as 'ban'.

The third feature of a gesture is the detailed temporal information, which concerns how gestures unfold over time. Figure 1.4 shows the five gestural landmarks (see also Gafos, 2002). These landmarks correspond to five key stages during the unfolding of a gesture: onset, target, maximum constriction, release and offset. The onset of movement is the point where the articulator begins moving towards the target. The target (or target achievement) is the point in time where the constriction attains the goal, known as the achievement of the target. After reaching the target, the constriction is held for some time. Then, the articulator begins releasing the constriction and moving away from the target. This point is the release. The offset is the end of the constriction. The interval between the achievement of the target and the release is referred to as the plateau.

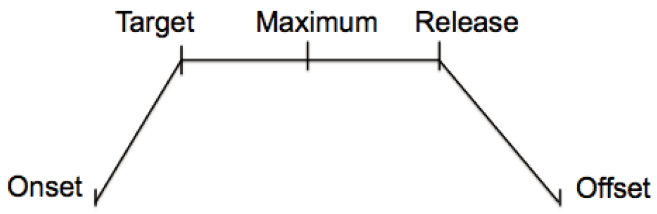


Figure 1.4: Gestural landmarks. See also Gafos (2002)

As mentioned above, gestures are 'glued' together to ensure the temporal stability of the information pattern. Gafos (2002) investigated how articulatory gestures are organized over time. For instance, in the English word 'act', there is no

audible release between the first consonant, /k/ and the second, /t/. In this case, the target of the second consonant (C2) is synchronous with the release of the first (C1), and the two consonants are produced in a temporal pattern called 'close transition'. A schematic representation of a close transition is shown in Figure 1.5. The interval between the plateaus, or inter-plateau interval (IPI) is zero.

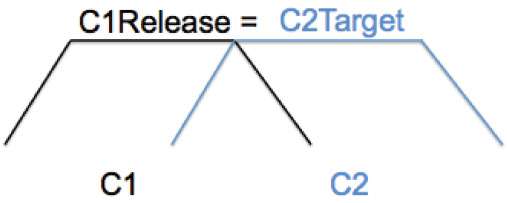


Figure 1.5: A high overlap gestural coordination pattern with a close transition between two consonants

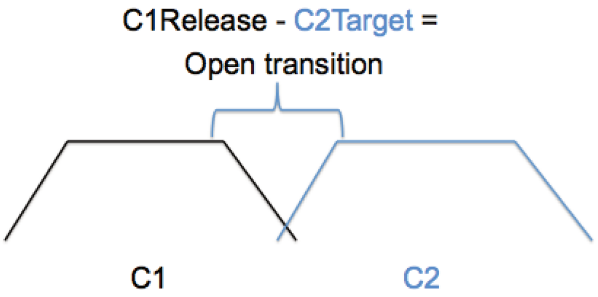


Figure 1.6: A low overlap gestural coordination pattern with an open transition between two consonants

When the gestures of the two consonants in C1C2 clusters have low overlap, this coordination pattern leaves a short period of time between the two consonants during which the vocal tract remains open, from the release of C1 to the target of C2. Acoustically, this open transition may give rise to an vocoid, most often a schwa-like sound (Catford, 1985). A schematic representation of an open transition

is shown in Figure 1.6. In this case, the inter-plateau interval (IPI) between the two consonants is greater than zero.

In an articulatory study of Moroccan Arabic, Gafos et al. (2010) showed that forms such as [katᵊb] 'to write' are produced with a schwa-like element between the two final consonants /t/ and /b/. Gafos et al. attribute this inserted vocalic element to insufficient overlap of the two flanking consonants. That is, the schwa-like vocoids do not have a gestural target, as lexical vowels do.

Cross-linguistically, temporal coordination patterns vary between 'close transition' (i.e. high overlap of two consonantal gestures) and 'open transition' (i.e. low overlap of two consonantal gestures). Zsiga (2000, 2003) studied temporal coordination in stop-stop clusters at word boundaries in English and Russian, measuring gestural overlap based on acoustic data. Her results show that English exhibits a greater degree of overlap than Russian, and a lower rate of C1 release. By contrast, C1 release was found to be obligatory in Russian, a constraint which results from a lower degree of overlap between consonants at word boundaries.

We have seen that consonant clusters can vary in their amount of overlap. Different temporal coordination patterns trigger different acoustic patterns. In the next subsection, we will review patterns of temporal overlap found in the production of non-native consonant clusters.

1.3.2 Temporal coordination in non-native CCs

In the production of non-native consonant clusters, speakers may repair illicit consonant clusters in different ways, which include inserting a vocoid between two consonants, deleting one of the consonants, or changing one of the consonants.

Davidson (2006b) argued that the nature of vocoids between consonants may be not a lexical vowel, but an open transition due to insufficient temporal overlap of the flanking consonants. She proposed an extension of Gafos's (2002) gestural

coordination model to the production of non-native clusters.

Figure 1.7 shows the two representations proposed by Davidson for gestural coordination resulting in an vocoid between two consonantal gestures (see also Hall, 2006). There are two possible scenarios:

- If the vocoid is a lexical vowel, this vowel should have its own vocalic gesture and its own gestural target.
- If the vocoid corresponds to a transition, then no extra segment is added: the transition is a by-product of the temporal coordination of the surrounding gestures.

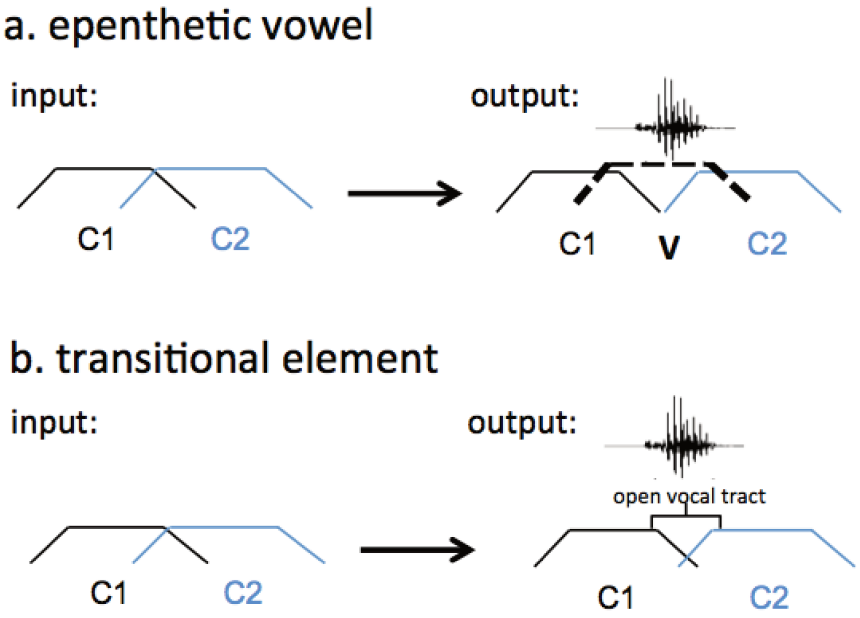


Figure 1.7: Gestural coordination with insertion between two consonants: (a) inserting a lexical vowel between two consonants; (b) inserting a transition due to gestural mistiming. See also Davidson, 2006a.

As reviewed in Section 1.2, previous studies regarding the perception of non-native consonant clusters showed that listeners may hear a vowel within non-native consonant clusters (Dupoux et al., 1999; Berent et al., 2007; Davidson and

Shaw, 2012). These vowels modify consonants clusters that are illicit in the native language. Thus, the inserted vowels have a phonological function, just like a lexical vowel. The production of such an inserted vowel is as shown in Figure 1.7 (a), with a vocalic gesture inserted between two consonantal gestures C1 and C2.

However, Davidson (2005, 2006a) suggested an alternative account based on gestural coordination. According to her, vowel insertion may be the result of failure to adequately overlap the gestures in the production of non-native clusters. This corresponds to the open transition (cf. Catford, 1985), the delay between the release of the first consonant and the target achievement of the second one. Unlike in the case of epenthesis, there is no vocalic gestural target between the two consonants, as shown in Figure 1.7 (b). Catford calls this configuration 'gestural mistiming'.

According to Davidson (2006a), 'gestural mistiming' is not due to lack of motor skill with non-native consonant clusters, but rather to the fact that the native phonology of English prohibits speakers from using the native timing pattern to produce non-native consonant clusters. Since English has a high overlap pattern, i.e., close coordination between consonants (see 1.5), native phonotactic knowledge prevents the use of close coordination for illegal sequences. Thus, consonant sequences that do not occur in English cannot be produced with the native close coordination timing pattern. They will be produced instead with 'gestural mistiming', resulting in less overlap between consonants.

In an experimental study, Davidson (2006a) compared the acoustic properties of the schwa-like vocoid inserted by English speakers to repair illicit clusters with those of lexical schwa /ə/ in English. By measuring the duration and formant frequencies (F1 and F2 midpoint) of the inserted vocoid, she sought to determine its nature as either a transitional or an epenthetic schwa. Davidson predicted that if the vocoid is a 'transition', its acoustic properties should be different from

those of lexical schwa, with lower F1 and F2, and a shorter duration. Since the vocoid, under this view, is a brief period of open 'transition' between the preceding consonant and the next, the oral cavity should be more closed than it is for lexical schwa, resulting in lower F1. Moreover, since the 'transition' lacks a specified target for schwa, the tongue root should anticipate the position of the first vowel following the cluster (in all cases /a/ in Davidson's experiment), potentially resulting in lower F2. If, on the contrary, the vocoid is a lexical vowel, its acoustic properties should depend on the surrounding consonants, with, for instance, lower F2 in labial and other front contexts, but higher F2 in dorsal contexts. As for F1, it should be equal to or greater than the F1 of a lexical schwa. Davidson found that English speakers most often modified illicit consonant clusters by inserting a vocoid, and that both F1 and F2 of the inserted vocoids were significantly shorter than they were for lexical vowels. The duration of the vocoids was also much shorter than that of a lexical vowel. These results led Davidson to conclude that the presence of such vocoids results in less overlap of the flanking consonants. In the current study, we used the same acoustic measurements as Davidson (2006a) to determine the quality of inserted vocoids. We will present the results in Section 4.3.2

We saw that less overlap of flanking gestures in the production of non-native CCs may trigger an acoustic 'transition' between two consonants. However, even in cases where no such 'transition' occurs, the L2 timing patterns may still be produced differently from the L2. To further examine this question, it is worth returning here to an important study by Zsiga (2003) that investigated L2 timing patterns. Zsiga tested native clusters in both L1 and L2. The stimuli used were stop-stop consonant clusters across word boundaries in English and Russian. The main question addressed by Zsiga's experiment was whether L2 production is transferred from the L1 timing pattern, or driven by universal phonetic factors—such as perceptual recoverability—where two consonantal gestures are less overlapped

so as to include sufficient information in the acoustic signal to facilitate correct perception (Wright, 1996; Chitoran et al., 2002; Kochetov, 2006). In English, the degree of overlap at word boundaries is considerably high, with a rarely audible release burst of C1 in C1 # C2 sequences (# indicates a word boundary). In contrast, Russian exhibits an obligatory audible release burst of C1 at word boundaries, which is the result of reduced temporal overlap, as shown by Zsiga (2000); Kochetov et al. (2007).

Results showed that native speakers of Russian used their native timing pattern, characterized by less overlap, in the production of English words. However, native English speakers did not similarly transfer their native pattern—more overlap—to L2 production. Instead, they used a similar low-overlap timing pattern across all clusters, which was neither English nor Russian. Zsiga argued that both English and Russian speakers tended to use the reduced overlap pattern in L2, so as to produce an audible release for listeners. Russian speakers maintained their native timing pattern, while English speakers could neither retain their native timing pattern nor reproduce the timing pattern of Russian. According to Zsiga, these results support a role for both transfer from L1 and the emergence of the unmarked in timing patterns of L2 speech.

As for L2 timing patterns in a language with moderately complex syllable structure, Yanagawa (2006) pioneered a study of timing patterns in several L1s including English, German, Japanese, and Cantonese, and in L2 English. She found that Japanese speakers exhibited a low degree of overlap in the production of consonant clusters in L1. However, it should be noted that since Japanese has very few consonant clusters, all the clusters tested in Yanagawa's study were CVC sequences with devoiced vowels. Thus, even though there were no vocoids occurring between the two consonants, the vocalic gestures may still not have been deleted. As regards gestural coordination in L2, Japanese speakers correctly reproduced

English consonant clusters and exhibited a similar timing pattern to that used by native speakers of English. That is, instead of producing CC clusters with a low degree of overlap, as in their L1, Japanese speakers exhibited a high degree of overlap in the production of English clusters. However, Yanagawa stressed that the results were based on a very small sample size, and further that the participants had good knowledge of L2 English. In particular, Yanagawa observed a frequency effect whereby the more familiar the Japanese speakers were with an English word, the greater the degree of overlap they produced for that word. For instance, *pt* in 'captain' was more overlapped than *pt* in the less frequent word 'cryptic'. This effect was also observed in the speech of English native speakers.

Luo (2017) examined the effect of word frequency on the production of English consonant clusters by native speakers of both English and Mandarin, testing stop-stop combinations across word boundaries (e.g., 'take care'). Consistently with Yanagawa's results discussed above, Luo found a significant correlation between the frequency of the English words (as reported by participants based on their own assessment) and the degree of overlap with which they were produced.

In the current study, the main goal is to investigate how a language's native timing pattern affects its speakers' production of non-native consonant clusters. To do so, we tested monolingual speakers of Mandarin who had never learned a second language, or had very low proficiency in their L2. Given that these speakers have had very little exposure to the tested consonant clusters, it is reasonable to hypothesize that they will perform differently from speakers of languages such as English, who have more experience with such clusters.

In the next subsection, we will present Mandarin Chinese phonology, focusing on syllable structure in Mandarin. Then, we will present our hypotheses on the perception and production of non-native consonant clusters by monolingual speakers of Mandarin.

1.4 The current study

1.4.1 Why Mandarin?

We have already pointed out that, when it comes to studying the adaptation of non-native clusters by speakers of a language with simple or moderately complex phonotactics, most earlier studies examined Japanese as a test language. There are comparatively fewer studies on Mandarin Chinese (Broselow and Finer, 1991; Wang, 1995; Broselow et al., 1998; Hansen, 2001; Fan, 2011; Luo, 2017; Durvasula et al., 2018), even though the phonotactics of Mandarin are very similar to those of Japanese. Like Korean and Japanese, Mandarin has a moderately complex syllable structure according to the WALS categorization (Dryer and Haspelmath, 2013, available online at http://wals.info/). The inventory of consonant clusters in Mandarin is very small. The only legal onset clusters in Mandarin are the consonant-glide combinations: C + /w, j, ɥ/. For this reason, we expect a vowel to be perceived or inserted by native speakers. Another similarity with Japanese (and with Korean) is the presence of devoiced vowels in Mandarin. According to Duanmu (2007), Mandarin non-low vowels may be devoiced after aspirated consonants in low tone contexts. Unlike Korean and Japanese, Mandarin has no voicing contrast in obstruents. The stop consonants are either voiceless non-aspirated, i.e., /p, t, k/ or voiceless aspirated /p^h, t^h, k^h/.

Most of the previous studies reviewed above tested the adaptation of non-native consonant clusters from English, which has a limited cluster inventory. In the current study, we test the adaptation of consonant clusters from Russian, a language with complex syllable structure according to the classification of Easterday (2017).

1.4.2 The phonemic inventory of Mandarin

Table 1.2 shows the consonantal inventory of Mandarin, which includes twenty-two consonants and three glides. The dental nasal [n] and the velar nasal [ŋ] are the only possible consonants in coda position. The three glides [w, j, ɥ] do not contrast with the corresponding high vowels [u, i, y]. They can, however, form a complex onset (C-Glide) with the preceding consonant. We will discuss this phenomenon in Section 1.4.3.

Table 1.2: Consonant inventory of Mandarin. Lin (2007)

	bilabial	labio-dental	dental	post-alveolar	alveolo-palatal	palatal	velar
stop	p p^h		t t^h				k k^h
fricative		f	s	ʂ	ɕ		x
affricate			ts ts^h	tʂ $tʂ^h$	tɕ $tɕ^h$		
nasal	m		n				ŋ
approximant	w ɥ			ɹ		j ɥ	w
lateral			l				

Table 1.3 shows the vowel inventory of Mandarin. Most authors argue that Mandarin has five basic vowels /i, y, u, ə, a/ (Chao, 1965; Duanmu, 2007; Lin, 2007). However, Lee and Zee (2003) analyzed the production of a native speaker of Mandarin and suggested based on the results, that Mandarin actually has six vowels, including a close-mid back unrounded vowel /ɤ/.

	Front	Central	Back
Close	i y		u
Close-mid			(ɤ)
Mid		ə	
Open	a		

Table 1.3: Vowel inventory of Mandarin.

Duanmu (2007) and Lin (2007) argued that the vowel /ɤ/ is not a phoneme,

but a phonetic variant (allophone) of the central vowel /ə/. The vowel /ɤ/ only occurs in syllable-final position in an open syllable, e.g., [lɤ] 乐 'happy', while the central schwa /ə/ occurs only in a closed syllable, e.g., [ləŋ] 冷 'cold'.

1.4.3 Mandarin syllable structure

Table 1.4 shows the possible segmental combinations in Mandarin Chinese syllables, based on Duanmu (2009, p. 97).

Table 1.4: Possible segmental combinations in Mandarin Chinese syllables. Based on Duanmu (2009)

Position	Phonemes	Notes
C	p, pʰ, t, tʰ, k, kʰ, ts, tsʰ, tʂ, tʂʰ, tɕ, tɕʰ ɕ, f, s, ʂ, ʐ, x, m, n, l	One of 21 Cs (never ŋ), or no C
G	j, w, ɥ	One of three Gs, or no G
V	i, y, u, ə, a	One of five Vs
X	i, u, n, ŋ	One of five Xs, or no X

The maximal syllable structure allowed in Mandarin is CGVX, where C is a consonant, G a prenucleus glide, V a vowel, and X is either a nasal (N) or an off-glide (the off-glide is considered as the second vowel of a diphthong in Duanmu (2007)). A consequence of these restrictions on maximal syllable structure is that the permitted consonant sequences in Mandarin are very limited, as shown in Table 1.5.

However, there is debate over the status of the on-glides /j, w, ɥ/ in onset CG. There are three main hypotheses, summed up schematically in Figure 1.8:

(a) G is part of the Final (Rime). Traditional analysis shows that syllables consist of an optional Initial (I), which is a syllable initial non-glide simple consonant (not a consonant cluster). The rest of the syllable is formed of the

Table 1.5: Attested consonant sequences in word-initial position and across word boundaries in Mandarin. C is a consonant, G is a prenuclear glide, and N a nasal.

Consonant clusters		Examples		
		IPA	Pinyin	Gloss
initial-CG	#Cj	[ljæn]	lian	联 'union'
	#Cw	[lwæn]	luan	卵 'egg'
	#Cɥ	[lɥee]	lüe	略 'omit'
medial-N#C	n # C	[ljæn # tʰɔŋ]	lian#tong	联通 'to connect'
	ŋ # C	[kaŋ # tʰjee]	gang#tie	钢铁 'iron'

Final (F), which is divided into an optional Medial vowel (M) and a Rime (R). The Medial is a prenuclear glide (/j/, /w/ or /ɥ/). The Rime consists of an obligatory Nuclear vowel (N) and an optional Ending (E), which may be the second half of a diphthong or one of the nasals /n, ŋ/ (Cheng, 1973; Lin, 1989).

(b) G is part of the Onset (O) rather than part of the Rime (R). Two different analyses have been proposed for this approach: Bao (1990) consider CG as forming an onset consonant cluster (Figure 1.8, b1); Duanmu (2007) proposes that onset CG combinations are to be considered a single segment instead of two separate ones. He argues that the glide is co-articulated tightly with the preceding consonant, and as such forms part of the onset as a secondary articulation. In support of this hypothesis, Duanmu provides the following arguments: first, as observed by Chao (1965), /sw/ sounds different in English (e.g., [sweɪ] 'sway') and Chinese (e.g., [swei] 'age'). Namely, CG in Chinese is more coarticulated than in English. Second, in CGV sequences,

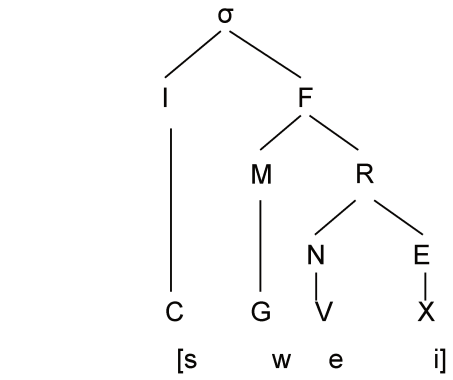
if CG does not have conflicting features, but GV does, CG can share the same time slot and V is in a separate slot. For instance, in the word /lia/ → [lʲaa] 'two', the features of /i/ and /a/ are in conflict. /i/ has features [+high, -low] and /a/ is [-high, +low] (see Duanmu, 2007). For this reason, /ia/ cannot occur in the same timing slot. Third, the duration of syllables with a prenucleus glide (CGVX) and without one (CVX) is almost the same. Fourth, the CG is sometimes replaced by C in casual speech, for example, /sj/ can alternate with /ɕ/.

(c) G is sometimes part of the Onset and sometimes part of the Rime. Yip (2003) analyzed the status of the prenuclear glide, and found G can be be part of the Onset, or part of the Rime, and can even be in both of them or in neither of them. Moreover, G is subject to both inter- and intra-speaker variation in Mandarin.

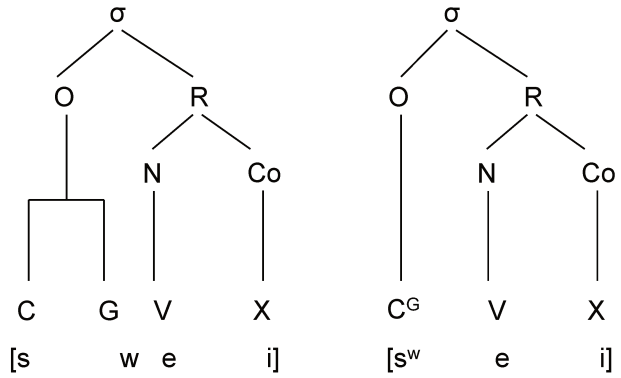
The debate on the status of the prenuclear glide in Mandarin is an ongoing one. However, since the choice between the above hypotheses is not directly relevant to the current study, it will not be addressed here. We are specifically interested in how unattested, non-native consonant clusters (stop-stop, stop-nasal, stop-liquid, liquid-stop, nasal-stop) are perceived and produced by native speakers of Mandarin, a language with moderately complex syllable structure.

1.4.4 Reduced vowels in Mandarin

We stated earlier that the maximum syllable structure in Mandarin is CGVX. Duanmu (2009) showed that most syllables in Mandarin have a full rime (VX), where VX may be a V+N sequence, a diphthong (VV), or a long vowel (V:), with an optional C onset. As Mandarin is a tone language, a lexical tone is obligatory for these syllables. Such syllables are called 'stressed syllables' or 'heavy syllables'. On



(a): Part of Final (Rime)



(b1): Onset cluster

(b2): Secondary articulation on Onset C

Figure 1.8: Diagram showing the syllable structure for different CG combinations. Where σ = syllable, O = onset, R = Rime, N = nuclear, Co = coda

the contrary, unstressed syllables usually bear a 'neutral tone' (and are therefore also described as toneless syllables). Such syllables have a reduced rime, meaning that the coda N is deleted, or the vowel is reduced. These syllables are usually a function word, or the second part of a compound word. Some examples of stressed and unstressed syllables in Mandarin are shown in Table 1.6.

Lin and Yan (1980) measured the duration and quality of unstressed vowels. They found that unstressed syllables were reduced 50% of the time, and that the quality of vowels in unstressed syllables was centralized to a schwa-like sound.

Table 1.6: Examples of the stressed and unstressed syllables in Mandarin (Duanmu, 2009)

Stressed (IPA)	Unstressed (IPA)	Gloss	Example
[la:]	[lə]	ASP	[mai-lə] 'buy-ASP (bought already)'
[t^{h}ou]	[t^{h}o]	head	[mu:-t^{h}o] 'wood-head (wood)'
[fãŋ]	[fə̃]	direction	[ti:-fə̃] 'land-direction (place)'

In many languages, short, unstressed high vowels and schwa can also be devoiced when they are adjacent to voiceless consonants, especially in fast speech. For instance, the three high vowels of Korean can be devoiced partially or completely between two voiceless stops (Jun et al., 1997). The same is true of Japanese (Hirose, 1971; Shaw and Kawahara, 2018). As in these languages, high vowels in Mandarin can also be devoiced. This is especially true of the high vowels [i, ɤ, u, y] when they occur: (1) after a voiceless/aspirated onset, including voiceless fricatives, aspirated stops, and aspirated affricates; (2) in an unstressed position; (3) with a low tone; (4) in spontaneous and fast speech (Duanmu, 2007; Lin, 2007). In addition, vowel devoicing can occur in syllables in all word positions, and the duration of the devoiced syllable does not change much compared to that of a voiced syllable (Duanmu, 2007).

Based on the existence of vowel devoicing in Mandarin, and on the perceptual repair of non-native consonant clusters by epenthesis, Zhao and Berent (2016) proposed that when Mandarin speakers hear a stop-C cluster, they perceive a long burst release followed by aspiration as a devoiced vowel in Mandarin. In the next subsection, we will present Mandarin inserted vowels in greater detail.

1.4.5 Inserted vowels in Mandarin

There is evidence that native speakers of Mandarin (like native speakers of Japanese) very often perceive a vowel between two consonants in non-native consonant clus-

ters, and insert an vocoid between the consonants when producing those clusters. This phenomenon has been reported in several domains, such as loanword adaptation or second language acquisition, and has been documented by a number of experimental studies (Miao, 2005; Guo and Nogita, 2013; Zhao and Berent, 2016; Durvasula and Kahng, 2015). However, unlike Japanese, which has only one inserted vowel to repair illicit consonant clusters (the close back unrounded vowel /ɯ/), Mandarin has different inserted vowel depending on the consonantal context. Miao (2005) analyzed a corpus of modern loanwords in Mandarin and observed that the quality of vowels between two consonants in consonant clusters agreed in place of articulation with the preceding consonant. For instance, a labial/rounded vowel /u/ occurs mainly after labial consonants, e.g., /p, m, f/, but a non-labial/unrounded vowel /ɤ/ tends to occur after non-labial plosive consonants, e.g., /t, k/. The high vowel /i/ occurs after palatal consonants, e.g., /ɕ, tɕ, tɕʰ/. The apical approximant /ɻ/ is inserted after the retroflex affricates /tʂ, tʂʰ/ and fricative /ʂ/. See the following examples from Miao (2005) (Table 1.7).

Table 1.7: Examples of inserted vowels in Mandarin loanwords (Miao, 2005)

vowels	source words			loanwords in Mandarin
/u/	'BobDog'	/bɒbdɒg/	→	/pa-pu-tou/
/ɤ/	'Target'	/ˈtɑːrgɪt/	→	/tai-tɕi-tʰɤ/
/i/	'Orange'	/ˈɔːrɪndʒ/	→	//ɑu-lan-tɕʰi/
/ɻ/	'Yuppies'	/ˈjʌpiz/	→	/ia-pʰi-ʂɻ/

Durvasula and Kahng (2015) argue that the quality of the vowel is determined not only by the place of articulation of the consonant, but also by the phonological patterns of Mandarin. They conducted a perception experiment with Mandarin speakers, and found that they only perceived the vowel [i] after the aspirated alveolo-palatal affricate [tɕʰ] in [atɕʰma] and only the schwa [ə] after the aspirated alveolar stop [tʰ] in [atʰma], since in Mandarin, [i] is not allowed after [tʰ] and [ə] is

not allowed after [tɕʰ]. Therefore, both phonological patterns and place of articulation of the preceding consonants influence the quality of vowels in Mandarin.

However, regarding production, Guo and Nogita (2013) analyzed acoustic data from Mandarin learners of English and found that the quality of inserted vocoids was a schwa-like mid-central articulated vocoid. They observed that such vocoids occurred after different English consonants, such as /k, g, b, ʃ/. This suggests that in production, the quality of such vocoids is not influenced by the place of articulation of the preceding consonants in production.

In the current study, we will report on the quality of inserted vowels (or vocoids) in Mandarin in both the production and perception of non-native consonant clusters.

In the final section of this introductory chapter, I will lay out the research questions that will be the focus of this dissertation, and the related hypotheses.

1.5 Research goals and hypotheses

The goal of this dissertation is to broaden our understanding of the perception and production of non-native phonotactics by studying how native speakers of Mandarin, a language with moderately complex syllable structure, perceive and produce consonant clusters. Relying on previous studies (Dupoux et al., 1999; Davidson, 2006a; Yanagawa, 2006; Wilson et al., 2014; Durvasula et al., 2018, among others), the overarching question of our study is the following: to what extent are the production and perception of non-native CC sequences influenced by phonological knowledge of the native language (phonotactic knowledge and gestural coordination patterns of the native language), and to what extent is it influenced by sensitivity to phonetic details (such as duration and intensity of the burst release)? Two specific hypotheses are tested with respect to Mandarin:

1. If native phonotactic knowledge alone is affecting non-native perception and production, then Mandarin speakers are expected to:

   (a) perceive illusory vowels systematically in consonant clusters, and

   (b) produce transitional vocoids systematically in such clusters.

2. If sensitivity to phonetic details primarily contributes to non-native perception, then Mandarin speakers are expected to:

   (a) report different types of perceptual modifications for different types of clusters, depending on their phonetic properties, and

   (b) produce different types of modifications for different types of clusters, when exposed to non-native stimuli.

In the current study, hypothesis (1a) and its alternative (2a) are tested in an ABX discrimination experiment, which will be presented in Chapter 2. This experiment tests Mandarin speakers' ability to discriminate CC from CVC sequences. If native phonotactic knowledge drives Mandarin speakers' perception, we predict that Mandarin participants will have difficulty discriminating CC clusters from CVC. We predict that Mandarin speakers will always perceive an illusory vowel in the clusters, making it difficult to distinguish them from CVC.

If, on the contrary, sensitivity to phonetic detail drives perception, Mandarin native listeners are predicted to have difficulty discriminating CC from CVC only for certain types of consonant combinations, depending on their phonetic properties. This prediction is based primarily on the results of Wilson et al. (2014) and Zhao and Berent (2016). Wilson et al. (2014) found that native speakers of English were sensitive to phonetic details present in the Russian stimuli. Specifically, shorter burst duration and/or lower amplitude of C1 led to more consonant deletion or to the perception of different consonants. Conversely, longer burst duration

and higher amplitude of C1 led to more frequent perception of vowels in the non-native clusters. Zhao and Berent (2016) also found that duration and intensity of C1 release affects the frequency of vowels perceived by Mandarin speakers. They proposed that Mandarin speakers perceived longer duration and higher amplitude of C1 release as a devoiced vowel following an aspirated consonant. Hearing a vowel was not the only perceptual modification of non-native clusters reported in Zhao and Berent's study. The authors, however, did not discuss whether the phonetic properties of the C1 release also influenced other perceptual modifications. In the current study we pursue this question further. Based on the results of Zhao and Berent (2016), we predict that Mandarin speakers are sensitive to the duration and intensity of C1 release, and that they may use such information in different ways for different types of clusters. To further investigate these details, we conducted a free-style transcription experiment in which participants were asked to write down the non-native stimuli that they heard. This experiment is presented in Chapter 3.

We hypothesized that if native phonotactic knowledge drives perception, native speakers of Mandarin will not only systematically perceive vowels in the non-native clusters (Hypothesis 1a), but they will also systematically produce them (Hypothesis 1b). However, if Mandarin speakers are sensitive to phonetic details, we expect that Mandarin speakers may also produce non-native CCs with other types of modifications (Hypothesis 2b). Hypothesis (1b) and its alternative (2b) are tested in a prompted production experiment, in which Mandarin native speakers heard stimuli containing CC and CVC sequences produced by a native speaker of Russian, and were asked to repeat what they heard. This experiment is reported in Chapter 4.

In the next chapter, we will investigate the effects of phonotactic knowledge and phonetic details (such as duration and intensity of the burst release) on the

perception of non-native consonant clusters by monolingual speakers of Mandarin using an ABX discrimination experiment.

Chapter 2

Perception of non-native CCs

The aim of the current section is to understand to what extent the perception of non-native CC sequences is influenced by phonotactic knowledge of the native language, and to what extent it is influenced by sensitivity to the following phonetic details: duration and intensity of the burst release.

A discrimination experiment was conducted to investigate these issues. It is designed as an ABX discrimination test in that listeners heard three stimuli per trial (labeled A, B and X) and reported whether X sounded the same as A or as B. For instance, in a trial such as CC_A - CVC_B - CVC_X, participants had to say whether CVC_X was the same as CC_A or as CVC_B. Two physically different tokens of the same stimulus were chosen for each ABX trial to encourage Mandarin speakers to respond only to phonetically relevant differences, rather than to any auditorily detectable differences. We predict that:

- If native phonotactic knowledge drives Mandarin speakers'perception, Mandarin participants will have difficulty discriminating CC clusters from CVC. We predict that Mandarin speakers will always perceive an illusory vowel in the cluster, making it difficult to distinguish them from CVC.

- If, on the contrary, sensitivity to phonetic detail drives perception, Mandarin native listeners are predicted to have difficulty discriminating CC from CVC only for certain types of consonant combinations, depending on their phonetic properties.

2.1 Methodology

2.1.1 Participants

The experiments were carried out in Beijing. 15 monolingual speakers of Mandarin participated in total, including five male and ten female speakers, aged from 32 to 55 (the mean age being 41). All participants were living in Beijing at the time of the experiment. In a language proficiency and background questionnaire, all participants reported that they had been born and raised in Northern China and had never lived in a different country or traveled abroad. All of them reported a low level of proficiency in English. They had learned English in high school, but none of them were currently studying or speaking English on a regular basis. None of them reported a history of speech or hearing impairments. All participants received information about the experiment before starting, signed consent forms, and received financial compensation for their participation after the experiment.

2.1.2 Stimuli

Table 2.1 shows the recorded auditory stimuli used in the experiment. The eight target stimuli consist of non-words containing CC clusters that are non-native to Mandarin Chinese. Half of the stimuli are of the form *C1C2'V1C3V2* and contain word-initial clusters; the other half are of the form *'V1C1C2V2* and contain clusters in word-medial position. In addition to the eight target stimuli containing four

Table 2.1: Stimuli used in the ABX discrimination test (/á/ denotes a stressed vowel)

	Non-native clusters		Native cluster
	Word-initial/pre-stress	Word-medial/post-stress	Word-medial/post-stress
Target stimuli(CC)	tkápa, ktápa, tpáka, ptáka	átka, ákta, átpa, ápta	ánka
Control stimuli (CVC)	tɐkápa, kɐtápa, tɐpáka, pɐtáka	átəka, ákəta, átəpa, ápəta	ánəka

different clusters in two word positions, the nasal-stop cluster /nk/ was also used in the non-word /ánka/. Since /nk/ is the only word-medial cluster attested in Mandarin, it may be taken as a reference, and considered a control cluster. We expect that Mandarin speakers are more likely to perceive a cluster in /ánka/ (/á/ denotes a stressed vowel) than in the target non-native clusters.

Control stimuli without CC clusters were also used. They all contained the same C1 and C2 as the clusters, separated by a vowel, and were of the form *C1V0C2'V1C3V2* or *'V1C1V0C2V2*. All stimuli were recorded by a female native speaker of Russian, a language in which all five CC clusters (/tk, kt, tp, pt, nk/) are attested. Thus, all stimuli were phonotactically possible words of Russian, with the exception of /tpáka/. The cluster /tp/ does not occur in word-initial position in Russian. However, the Russian speaker was able to produce this word naturally, without inserting an vocoid within the cluster.

Regarding word-level prosody, word-initial CC sequences were all in pre-stress position, preceding the main stress, while CC sequences in medial position always followed the main stress. In the CVC control words, the vowel between C1 and C2 was always an unstressed /a/, and therefore produced as a reduced, centralized vowel (Hamilton, 1980). In pre-stressed syllables, the surface realization of unstressed /a/ in Russian is reported to be a low mid central [ɐ], and in post-stressed and unstressed syllables, it is a mid central [ə] (Padgett and Tabain, 2005). In the current study, we measured the acoustic properties of the unstressed

vowels, including duration and F1/F2 midpoint values. In pre-stressed syllables (in word-initial position), the average duration was 58ms (SD = 4), the average F1 was 774 Hz (SD = 51), and the average F2 was 1579 Hz (SD = 82). In post-stressed syllables (in word-medial position), the average duration was 57ms (SD = 9), the average F1 was 696 Hz (SD = 51), that of F2 was 1636 Hz (SD = 70). These values are consistent with Hamilton (1980) and Padgett and Tabain (2005), confirming that the vowel in CVC controls was a reduced and centralized vowel. We will therefore use the symbol [ɐ] to denote pre-stressed vowels (e.g., [pɐtáka]), and [ə] to denote post-stressed vowels (e.g., [ápəta]), as well as unstressed vowels in final position.

The Russian native speaker who recorded the stimuli was a graduate student who had been living in France for one year and a half at the time of the recording. The speaker reported that she was highly proficient in French, but continued to speak Russian regularly (at least two hours per day). The recordings were carried out in a sound-treated recording booth. The stimuli were digitally recorded as .wav files via the Praat program (Boersma and Weenink, 2014) onto a MacBook Air laptop, using a Roland UA-55 Quad Capture USB audio interface and a RØDE NT1-A microphone, at a sampling rate of 44.1 kHz. The microphone was positioned approximately 40 cm away from the mouth of the speaker. Non-word stimuli were inserted into the following carrier sentence (where _ indicates the position of the non-word): Наш _ плывёт по морю [naʂ _ plɨvʲˈɵt pə ˈmorʲu] 'Our_is swimming in the sea'. All sentences were written in Cyrillic and presented on a computer screen one at a time using PowerPoint. Each single sentence appeared randomly eight times in total over the entire recording session, so that eight repetitions were recorded for each stimulus. The Russian speaker was asked to read each sentence one by one, naturally. She was instructed to make sure she pronounced the non-words exactly as if they were real Russian words.

All stimuli were normalized using the 'scale intensity' function in Praat to obtain a mean intensity of 70 dB. A phonetician and native speaker of Russian listened to the recordings and judged whether each stimulus had been correctly pronounced. All tokens were approved. To encourage participants to respond only to CC and CVC differences, rather than to any other auditorily detectable differences, two repetitions out of the eight recorded were selected to be used as stimuli in the perception experiment. The selection was based on the acoustic similarity between the repetitions, as evaluated by the experimenter. Thus, for each target cluster, four tokens were retained for the experiment. Figure 2.1 shows the spectrograms of the two repetitions of the target CC stimulus /ptáka/, and the two repetitions of its CVC control /patáka/. The /áka/ portion had similar acoustic properties in the four selected tokens with respect to vowel duration, formant frequencies (F1 and F2) and pitch (F0). The segmentation of the stimuli was determined by means of both auditory and visual inspection in Praat.

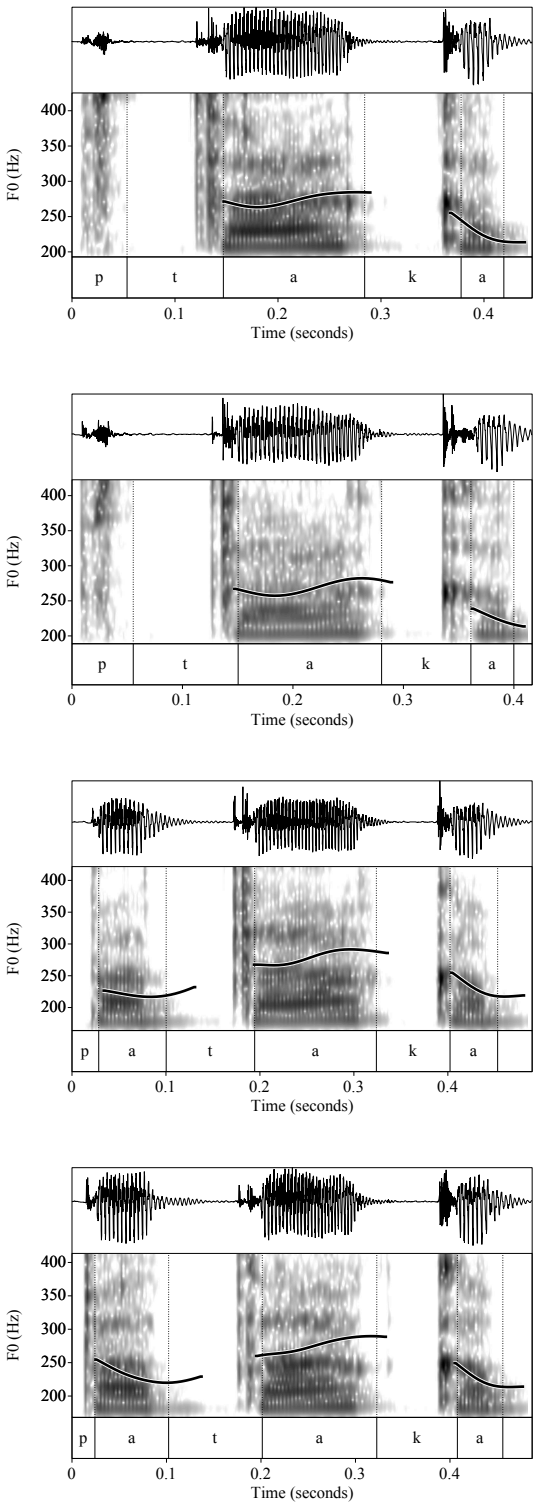
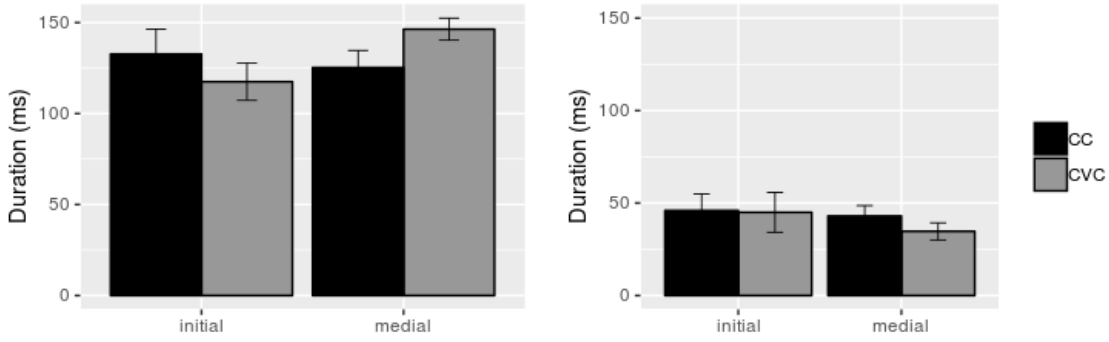


Figure 2.1: Spectrograms and waveforms of the two repetitions of /ptáka/ and its control /patáka/. F0 curves were plotted.

Figure 2.2 shows the duration of the vowels following the consonants of the cluster (V1), and of the word-final vowels (V2) in target CCs and CVC controls, across word-initial and word-medial positions. Error bars represent standard deviations. The figure indicates that the durations of these vowels are very similar in target CCs and CVC controls. Two linear models, one for V1 and the other one for V2, using the *lme4* package (Bates et al., 2014) in R (R Core Team, 2018), indicate that there was no significant difference in either V1 or V2 duration across target CCs and CVC controls (V1 duration: $\beta = 3.37$, $t(32) = 0.65$, $p = 0.52$; V2 duration: $\beta =$ -4.37, $t(32) =$ -1.4, $p = 0.16$, respectively).

Figure 2.3 shows the F1 and F2 midpoint values (Bark) of V1 and V2 in target CC and CVC control across word-initial and -medial positions. Error bars represent standard deviations. The F1 and F2 midpoint values (Bark) of these vowels are very similar in target CCs and CVC controls. Two multivariate regression models, one for V1 and the other for V2, using the MCMCglmm package (Hadfield et al., 2010) in R, show that there was no significant difference across the two conditions—target CC and CVC control—between F1 and F2 of either the first or the second vowels. (V1: *F1* = 0.13, *95% HPD* [-0.18, 0.44], $p = 0.41$; *F2* = 0.1, *95% HPD* [-0.26, 0.48], $p = 0.58$; V2: *F1* = 0.06, *95% HPD* [-0.31, 0.47], $p = 0.75$; *F2* = -0.07, *95% HPD* [-0.68, 0.48], $p = 0.84$, respectively).

Regarding pitch, given that Mandarin is a tone language, Mandarin speakers may perceive tones on vowels. We therefore asked a native speaker of Mandarin to listen to the stimuli and write down the tones she perceived. In the Russian stimuli, the stressed syllable is realized with an F0 peak, as shown in Figure 2.1. Thus, in both the target CC clusters and the CVC controls, the listener consistently perceived a low F0 in pre-stressed syllables, and a falling F0 in post-stressed syllables, respectively. In stressed syllables, she always perceived a high tone.



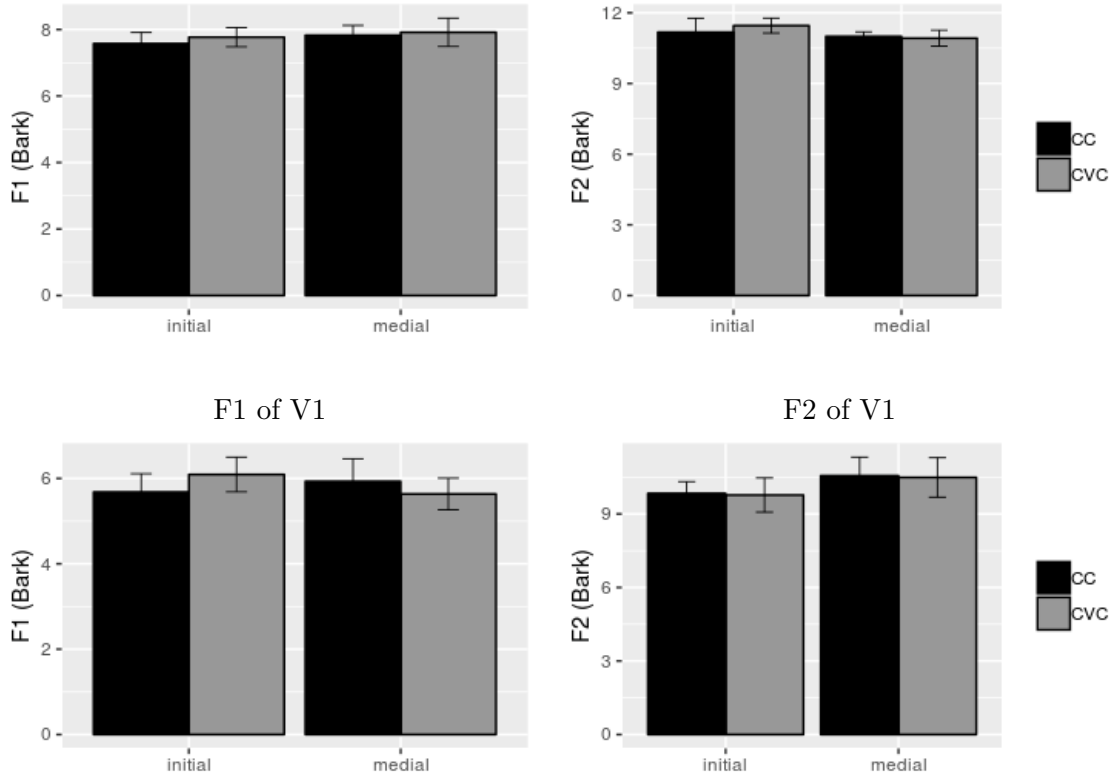
V1 V2

Figure 2.2: Duration (ms) of first vowels (V1), final vowels (V2) in target clusters CC (*C1C2**V1**C3**V2**/**V1**C1C2**V2***) and control sequences CVC (*C1V0C2**V1**C3**V2**/**V1**C1V0C2**V2***) across word-initial and -medial positions. Error bars represent standard deviations.

2.1.3 Procedure

The experiment contained 288 trials in total, and was divided into three experimental blocks with a self-timed break between blocks. Four practice trials were presented at the beginning of the experiment. The tokens used in the practice trials were different clusters from those in the experimental trials. Stimuli were differently randomized for each participant. Both response accuracy and response time were measured and analyzed. Participants did not receive any feedback during the test. The entire session lasted about fifteen to twenty minutes.

Each CC/CVC test pair was presented in four stimulus orders: ABA, ABB, BAA, BAB. For each ABX trial, two separate repetitions of the same stimulus were chosen. For instance, in a trial such as [ptaka]–[pɐtaka]–[pɐtaka], the first and third stimuli are different, and the correct answer would be to say that the third stimulus is the same as the second. A relatively long inter-stimulus interval of 1500ms was used to encourage participants to maintain the items in phonological working memory, a process that requires access to an item's phonological structure



F1 of V1 F2 of V1

F1 of V2 F2 of V2

Figure 2.3: F1 and F2 (Bark) of of first vowels (V1), final vowels (V2) in target clusters CC (*C1C2**V1**C3**V2**/**V1**C1C2**V2***) and CVC control sequences (*C1V0C2**V1**C3**V2**/**V1**C1V0C2**V2***) across word-initial and -medial positions. Error bars represent standard deviations.

so as to mentally 'assemble' that structure. Participants were instructed to respond using the response pad as quickly and accurately as possible after they hear the last word of the triplet.

Participants were seated in front of a computer equipped with a Cedrus RB-740 response pad. The experiment was run on Psychopy 2.0. The stimuli were played through Sony MDR-EX110AP in-ear headphones. Participants heard three stimuli per trial, and after each trial reported whether X was the same as A or as B. The instructions, given in Mandarin, are shown below with the English translation:

"在接下来的实验中，您将听到一些外语单词。您一次将听到三个词。最后一个词，有时和第一个词一样，有时和第二个词一样。所以您的任务是判断最后一个词和第一个词一样还是和第二个一样。如果您觉如果最后一个词和第一个词更相似，请按"红色按钮"，如果最后一个词和第二个词更相似，请按"蓝色按钮"。听完最后一个词后，请您以最快速度，做出最准确的选择。按空格键继续。"

"*In the following task, you will hear words in a foreign language. You will hear three words in a row. In some cases, the last word will be the same as the first. In others, the last word will be the same as the second. Your task is to decide whether the last word is the same as the first or the second. If you think the last word is the same as the first word, press the red button. If you think the last word is the same as the second word, press the blue button. Please respond as quickly and accurately as possible after you hear the last word.*"

2.1.4 Analyses

2.1.4.1 Sensitivity

To determine whether participants were sensitive to the presence or absence of a vowel between two consonants, we employed a signal detection theory (SDT) framework, which is commonly used to understand decision-making and accuracy in the presence of uncertainty (Green and Swets, 1966; Macmillan and Creelman, 2005). The d' measure, in particular, is widely used in speech perception (Iverson and Kuhl, 1995; Berent et al., 2009; Davidson and Shaw, 2012; Pouplier and Goldstein, 2005). During experiments, participants may not be absolutely certain about what they heard, potentially leading them to 'guess' the correct answer in some cases. To avoid this bias, SDT provides a statistical measure of partici-

pants' sensitivity to the difference between two signals (e.g., [ptaka] and [pʊtaka]). SDT uses two parameters, the sensitivity index (d') and response bias, to analyze participants' discrimination performance.

The current study involves an ABX discrimination task. We will only report on the sensitivity index d' (Macmillan and Creelman, 2005). d' was calculated based on the number of hits (correct matches of X to the A or B stimulus), and false alarms (incorrect matches of X to the A or B stimulus). Table 2.2 shows a simple response matrix.

	Response A	Response B
X matches A (AB-A or BA-A)	HIT A	FALSE ALARM B
X matches B (BA-B or AB-B)	FALSE ALARM A	HIT B

Table 2.2: A simple SDT response matrix for an ABX discrimination task

d' can be estimated from the observed rate of hits (H) and false alarms (F) using the formula

$$d' = z(H) - z(F)$$

where z(H) and z(F) are the z-transforms of the hit and false alarm rates, respectively.

However, if H (hit) = 1, or F (false alarm) = 0 (as is sometimes the case in our experiments), z cannot be calculated. In such cases, the number of hits and false alarms needs to be adjusted by adding or subtracting 0.5 from the frequency matrix (Kadlec, 1999; Macmillan, 2002).

In addition, we must take into account the different decision strategies that participants may use in this task. To determine which of two tokens A and B a third token X most closely resembles, there are two basic strategies discussed in Macmillan and Creelman (2005) and Hautus and Meng (2002): the independent-observation decision rule, and the differencing model. The independent-observation strategy

requires making two decisions: one on the order of the first two stimuli (AB or BA) and one on the value of the third (X). For example, in the current study, if stimuli A */ptaka/* and B */pɐtaka*] differed in [ɐ] insertion, participants would be interested in this particular difference, namely that there is no vowel between the consonants of the initial cluster in A (*ptaka*), whereas there is one in B (*p*[ɐ]taka). Participants would then consider the stimulus X *ptaka* independently, and decide whether a vowel was inserted in the initial cluster. If they decide that no such vowel occurs in *ptaka*, they will identify X as matching A, not B. In the differencing model, participants make only one decision: whether X is closer to A or to B. Hautus and Meng (2002) argued that participants in ABX tasks are more likely to follow the differencing model since it involves less cognitive effort, and because untrained participants may not have considered the alternative decision strategy (the independent-observation decision). Therefore, in the current study, we will report the *d'* scores based on both strategies—the independent-observation model, and the differencing model.

Any *d'* score above zero means that participants can perceive the difference between A and B above chance, with a higher *d'* indicating greater discrimination ability. The highest *d'* score is 4.65, which indicates that listeners perfectly discriminate between stimuli. On the contrary, a negative d' score demonstrates that listeners are not sensitive to the difference at all. The scores were calculated using a script developed by Christophe Pallier [1]. For each subject, the *d'* values were calculated across word positions (initial vs. medial) for each cluster type (/tk, kt, tp, pt, nk/). Each subject thus contributes one *d'* value per cluster type for both positions.

[1] http://www.pallier.org/computing-discriminability-a-d-and-bias-with-r.html

2.1.4.2 Response Time

Response time was measured only for the trials where participants gave correct responses. Response times were measured from the end of the last stimulus (X) in the ABX trial to the moment when participants pressed one of the buttons on the response pad.

2.1.4.3 Acoustic analyses of the Russian stimuli

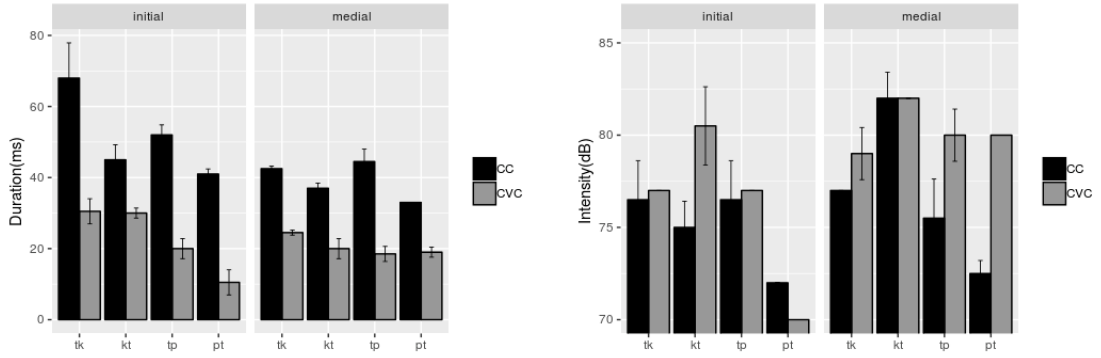
As shown in Figure 2.4, we analyzed the Russian stimuli to determine which acoustic properties may have affected Mandarin listeners' perception. We measured the stop burst duration and intensity of the first stop since, as noted earlier, Zhao and Berent (2016) suggest that Mandarin speakers are sensitive to such phonetic properties of the consonant clusters. Moreover, in Russian, burst releases are obligatory for stops (Zsiga, 2000) and differ in energy across different places of articulation (Stevens, 1998). Therefore, we included this measurement in the current analysis. All measurements were made in Praat. The dynamic range for spectrogram display in Praat was set to the default value of 70 dB. Burst duration of C1 was defined as the interval from the onset of burst to the onset of silence (onset of C2 constriction) on the waveform and spectrogram, while burst intensity was the maximum burst intensity during the C1 burst interval.

2.1.5 Statistical analysis

2.1.5.1 Sensitivity

Statistical analysis was carried out to evaluate whether participants' sensitivity to the difference between non-native CC clusters and CVC sequences was above chance. One-sample *t*-tests against $d' = 0$ (at chance) for each of these measures were carried out in R.

Figure 2.4: Duration (a) and intensity (b) of C1 burst of Russian stimuli for non-native CCs clusters and CVC controls sequences per cluster type across word-initial and -medial positions. Error bars represent standard deviations.



a:Duration of C1 release

b:Intensity of C1 release

In addition, to determine whether participants were more sensitive to the native medial cluster *nk* than to the non-native clusters, linear mixed-effect models were computed using the *lme4* package (Bates et al., 2014). The dependent variable was *d*' SCORE; the independent factors were CLUSTER TYPE across word-initial and word-medial positions (i.e., initial-/tk/, medial-/tk/, initial-/kt/, medial- /kt/, initial-/tp/, medial-/tp/, initial-/pt/, medial-/pt/, medial-/nk/). For contrast coding, we used dummy coding to compare each level of CLUSTER TYPE to a reference level. The reference level in the discrimination experiment was the native medial cluster /nk/. By-subject and by-item random intercepts were included. Random slopes for CLUSTER TYPE were not included, because the models including them failed to converge. *p*-values were derived using the lmerTest package (Kuznetsova et al., 2017) [2]. Since two statistical analyses were performed on the same sample of data, the Bonferroni correction was applied to *p*-values to adjust the familywise type I error rate. Therefore, a Bonferroni-corrected level of 0.025 (0.05 divided by two statistical analyses) was used for statistical sig-

[2]This was done using a Satterthwaite approximation to the degrees of freedom.

nificance.

2.1.5.2 Response time

Since the sensitivity scores (d') already take bias into account, we cannot use them to analyze the effects of response time on accuracy. We therefore computed a mixed binomial logistic regression model with accuracy (wrong–0 or correct–1) as the dependent measure. The independent variable was REACTION TIME (RT). By-subject, by-item random intercepts and random slopes for REACTION TIME were included.

Moreover, a separate model was conducted to test the effect of phonotactic knowledge: if the native cluster /nk/ could be recognized faster than the non-native clusters. The dependent variable was the correct RT, and the independent variable was CLUSTER TYPE (i.e., initial-/tk/, medial-/tk/, initial-/kt/, medial-/kt/, initial-/tp/, medial-/tp/, initial-/pt/, medial-/pt/, medial-/nk/). Dummy coding was used with the reference level medial-/nk/. By-subject and by-item random intercepts were included. Random slopes for CLUSTER TYPE were not included, because the model including them failed to converge. *P*-values were derived using the lmerTest package.

2.1.5.3 Acoustic analysis of the Russian stimuli

To determine the effect of C1 burst release on the discrimination between CC and CVC, we tested the results of the experiment against several linear mixed-effect models. The native *nk* clusters were removed from this analysis, since C1 is a nasal consonant. The dependent variable was *d*' SCORE. The independent variables were the acoustic properties of the Russian stimuli, including C1 BURST DURATION, C1 BURST INTENSITY and the consonant-to-consonant acoustic timing lags: IPI and IBI. The model also included random intercepts for participants and

clusters, and by-participant random slopes for the relevant independent variables. The acoustic measurements were nested in items, thus, no random effects for the items were included. *P*-values were derived using the lmerTest package.

2.2 Results

2.2.1 Sensitivity

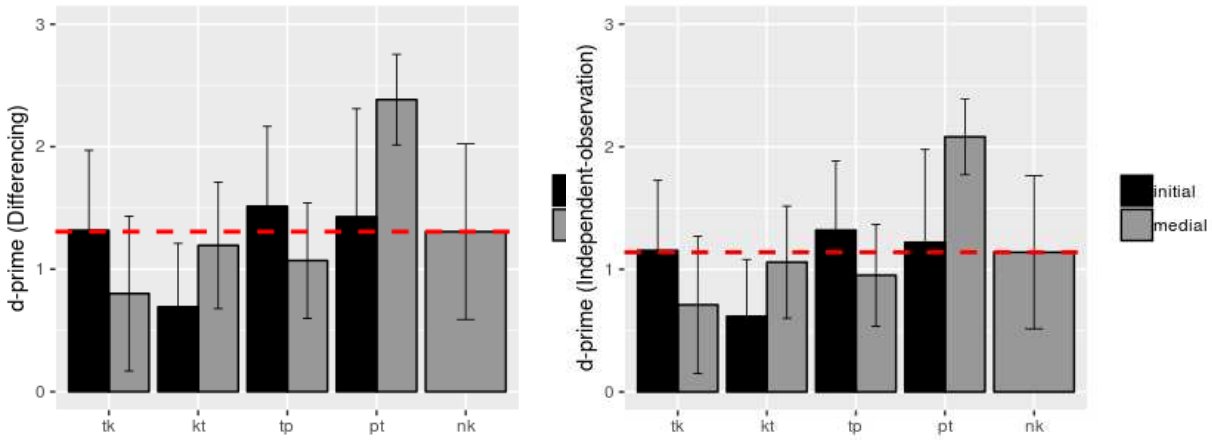
The results revealed that Mandarin speakers perceived the difference between non-native CCs and the CVC controls 65% of the time, which was significantly above chance (50% accuracy) ($t(1920) = 14.25, p < 2.2e-16, d$[3] $= 1.38$).

Table 2.3 summarizes the sensitivity scores of Mandarin speakers, based on the two discrimination strategies mentioned earlier, namely (a) the differencing strategy, and (b) the independent-observation decision strategy. Overall, *d'* was fairly high, indicating that Mandarin speakers were significantly sensitive to the contrast between CC and CVC. T-tests showed that *d'* was significantly above the chance score (zero) for both discrimination strategies. (Differencing strategy: $t(1920) = 44.43, p < 2.2e-16, d = 0.95$; independent-observation decision strategy:$t(1920) = 44.67, p < 2.2e-16, d = 0.95$).

Discrimination strategies	Sensitivity score (*d'*)			
	N	Mean	Std. Dev.	Range
Differencing	1920	1.3	1.38	−2.14 – 3.82
Independent-observation	1920	1.14	1.98	−1.89 – 3.24

Table 2.3: Summary of descriptive statistics of sensitivity scores (*d'*). *Note:* If *d'* > 0, participants were sensitive to the difference between CC and CVC. If *d'* < 0, participants were not sensitive to the difference. If *d'* = 0, the performance was at chance.)

[3] *d* is Cohen's d for the measure of effect size.



(a) Differencing (b) Independent-observation

Figure 2.5: Sensitivity (d') of Mandarin speakers to non-native CCs and native CVC sequences in ABX discrimination across word positions (initial vs. medial) for each consonant cluster, based on the differencing strategy (panel a) and the independent-observation strategy (panel b). Medial-*nk* (dashed red line), a licit cluster in Mandarin, was used as a baseline for comparison with other clusters. Error bars indicate 95% bootstrap confidence intervals.

Figure 2.5 shows the sensitivity (*d'*) of Mandarin speakers to non-native CCs and native CVC sequences across word positions (initial vs. medial) per consonant cluster, based on the differencing strategy (panel a) and the independent-observation strategy (panel b). The medial-/nk/ cluster, a licit cluster in Mandarin, was added as a baseline (shown in the figure by a dashed red line) to compare with the others. Error bars indicated 95% bootstrap confidence intervals. A linear mixed-effect model on sensitivity indicated that speakers were more sensitive to the difference between the native medial-/nk/ cluster, and its corresponding control /nək/ than the following clusters: medial-/tk/ vs. /tək/, medial-/tp/ vs. /təp/, initial-/kt/ vs. /kɐt/, for differencing strategy.(β = -0.51, $t(2145)$ = -5.08, $p <$ 4.19e-07; β = -0.24, $t(2145)$ = -2.37, p = 0.0179; β = -0.61, $t(2145)$ = -6.17, $p <$ 7.95e-10, respectively) For Independent-observation strategy, speakers were more sensitive to the difference between the native medial-/nk/ cluster, and its

corresponding control /nək/ than medial-/tk/ vs. /tək/ and initial-/kt/ vs. /kɐt/, but not medial-/tp/ vs. /təp/ (β = -0.43,$t(2145)$ = -4.93, $p <$ 8.79e-07; β = -0.52, $t(2145)$ = -6.01, $p <$ 2.12e-09, respectively). Participants were less sensitive to the CC vs. CVC contrast for medial-/nk/ than for medial-/pt/ for both strategies (Differencing: β = 1.08, $t(2145)$ = 10.82, $p <$ 2e-16; Independent-observation: β = 0.94, $t(2145)$ = 10.85, $p <$ 2e-16, respectively). For both strategies, medial-/kt/, initial-/pt/, initial-/tp/ and initial-/tk/, sensitivity to the CC vs. CVC contrast was not significantly different from the native medial-/nk/.

2.2.2 Response time

Figure 2.6 plots particpants' response times in the ABX discrimination experiment. The linear mixed-effect model used to assess the difference between the non-native clusters and the native cluster /nk/ indicated that responses for the native /nk/ cluster were significantly faster than for medial-/kt/ (β = 9.559e-02, $t(1390)$ = 2.085, p = 0.0372). Otherwise, there was no significant difference in response time between the native /nk/ and the other non-native clusters.

To assess the effects of correct response time on accuracy, we fit a mixed binomial logistic regression model. Results indicated that there was a significant effect of response time on accuracy to the CC and CVC contrast (β = -0.699, SE = 0.21, p = 0.000679). The estimated value of response time (β = -0.699, SE = 0.21) indicates that as accuracy of the contrast declined, response time increased. That is, the less accurate the contrast between CC and CVC, the longer it took for the decision to be made.

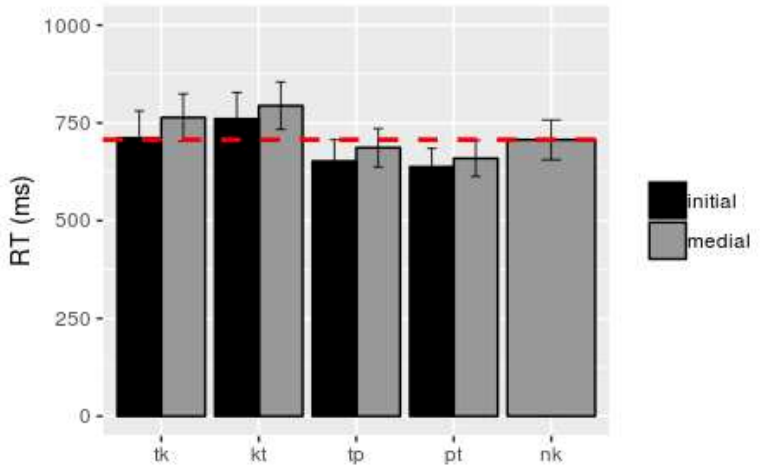


Figure 2.6: Response time of Mandarin speakers to non-native CC and native CVC sequences in ABX discrimination across word positions (initial vs. medial) for each consonant cluster. Medial-*nk* (dashed red line), a licit cluster in Mandarin, was used as a baseline for comparison with other clusters. Error bars indicate 95% bootstrap confidence intervals.

2.2.3 Acoustic analysis of the Russian stimuli

Our data showed that Mandarin participants were sensitive to the contrast between non-native CC and control CVC sequences, but there was a significant difference across cluster types. As shown in Figure 2.5, participants were significantly more sensitive to the contrast for the medial-/pt/ cluster than for others. The worst-perceived cluster in our data was initial-/kt/. Why is that? To answer this question, we analyzed the acoustic details of the Russian stimuli.

The results indicate that C1 burst intensity had a significant effect on the discrimination of non-native CC and control CVC sequences for both differencing and independent-observation strategies (Differencing: β = 6.344e-02, $t(1234)$ = 4.473, p = 8.48e-06; Independent-observation: β = 5.449e-02, $t(1234)$ = 4.411, p = 1.12e-05). When the burst release of C1 had higher amplitude, participants were less sensitive to the difference between C1C2 and C1VC2. For instance, for

medial-/pt/, which had a high sensitivity score (d' = 2.38, SD = 0.81), the burst intensity of the two tokens was 73 dB and 72 dB, respectively, which was lower than for the clusters with the lowest sensitivity scores, e.g., initial-/kt/ (d' = 0.69, SD = 1.14; intensity = 76 dB and 74 dB). In Russian, burst releases are obligatory for stops (Zsiga, 2000) and differ in energy across different places of articulation. In general, the energy of the air explosion at the release of a voiceless bilabial plosive /p/ is weaker than for the velar /k/ and the alveolar /t/ plosives (Stevens, 1998).

Higher burst amplitude results in lower sensitivity to the contrast between CC and CVC. As Zhao and Berent (2016) point out, Mandarin can have devoiced vowels following an aspirated consonant. Thus, native speakers may perceive a long burst release as an aspirated consonant followed by a devoiced vowel, because the burst has similar aperiodic energy to that of aspiration. Our results are partially consistent with those of Zhao and Berent (2016), who found that Mandarin speakers were highly sensitive to both C1 burst duration and burst intensity. Longer duration and higher amplitude of the burst of the first consonant triggered more epenthesis in the perception of non-native consonant clusters. However, in the current study, we only found an effect of stop burst intensity, not stop burst duration, on the perception of non-native consonant clusters. The duration of the C1 release burst did not affect discrimination of CC and CVC.

2.3 Interim summary

In this chapter, we studied the effect of phonotactic knowledge and phonetic details (specifially, C1 burst release) on the perception of non-native consonant clusters by Monolingual speakers of Mandarin. The results of an ABX discrimination task showed that Mandarin speakers were clearly sensitive to the contrast between

non-native CCs and native CVC control sequences. The *d'* score overall was significantly above chance for both the independent-observation and the differencing analysis. Moreover, we examined how participants' sensitivity varied between native and non-native consonant clusters. We found that they were significantly more sensitive to the native medial cluster /nk/ than to the non-native clusters medial-/tk/, medial-/tp/ and initial-/kt/ for the differencing strategy. As regards the independent-observation strategy, participants were more sensitive to the native medial cluster /nk/ than to medial-/tk/, and initial-/kt/, but not to medial-/tp/. For both strategies, participants were significantly less sensitive to medial-/nk/ than to non-native clusters medial-/pt/. In addition, there was a negative correlation between sensitivity and correct response time. That is, the less sensitive participants were to the difference between clusters and CVC sequences, the slower they were to make a decision about the CC vs. CVC contrast.

Sensitivity to non-native clusters was found to vary across cluster types, which suggests that certain phonetic cues may be more salient for Mandarin speakers, helping them to distinguish between CC and CVC. We observed that C1 burst intensity had a significant effect on the discrimination of non-native CCs and CVC control sequences. The higher the amplitude of C1 burst release, the less sensitive participants were to the difference between CC and CVC.

Chapter 3

Transcription of non-native CCs

From the ABX discrimination task presented in the previous chapter, we learned that monolingual speakers of Mandarin were able to distinguish CC from CVC correctly 65% of the time, even though all but one of the tested CCs (/nk/) were illicit in Mandarin. Their sensitivity to the CC–CVC contrast was significantly above chance, a result which is partly explained by the fact that speakers are sensitive to C1 burst intensity. Previous studies (Wilson et al., 2014; Davidson et al., 2015) proposed that higher amplitude of C1 is more likely to result in peceptual epenthesis, whereas lower amplitude of C1 can lead listeners to perceptually delete or change a consonant. Thus, non-native listeners might be perceiving different alterations of the stimulus depending on the amplitude, rather than systematically perceiving a vowel. These other types of misperceptions are not predicted to affect discrimination between CC and CVC by Mandarin speakers. To further test the extent to which speakers/listeners respond to phonetic details as opposed to phonotactic knowledge in the perception of non-native clusters, we conducted a transcription experiment, in which monolingual speakers of Mandarin were asked to write down the non-native stimuli that they heard using Pinyin (the official romanization system for Mandarin Chinese). It is predicted that:

- if native phonotactic knowledge primarily contributes to non-native perception, then Mandarin participants will transcribe all clusters with a vowel between C1 and C2;

- if sensitivity to phonetic details primarily contributes to non-native perception, then Mandarin participants will transcribe different types of clusters with different types of modifications, and the nature of these modifications will depend on the phonetic properties of each cluster type.

3.1 Methodology

3.1.1 Participants

The participants were twelve female and nine male monolingual Mandarin speakers aged between 28 and 52 (the average age being 38). Among them, six participants took part in the ABX discrimination task two years later. None of them reported a history of speech or hearing impairments. All participants were informed of the content of the experiment, signed consent forms before the experiment, and received financial compensation for their participation.

At the time of the transcription experiment, all participants were living in Beijing. In a language proficiency and background questionnaire, all reported that they had been born and raised in Northern China and had never lived or traveled abroad. Eleven speakers reported that they had never learned a second language, while the remaining ten reported a low level of proficiency in English. For these participants, an oral English test was conducted, asking them to read a list of English words or phrases in isolation, containing clusters in both word-initial and across-word positions: *stop, please, thanks, sit down, smoke, club.*

Figure 3.1 shows the frequency of the possible outputs for each of the six stimu-

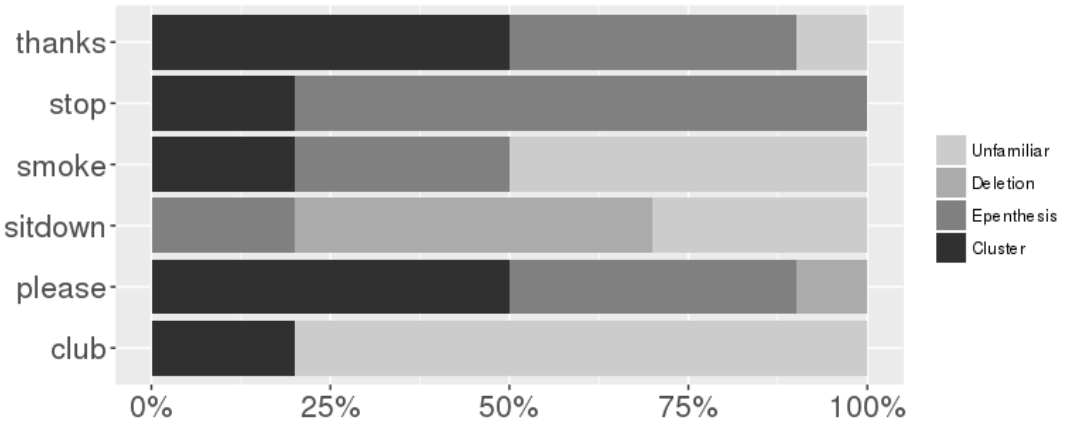


Figure 3.1: Outputs for the production of English words (unfamiliar words were not pronounced)

lus words, either correct—with a consonant cluster—or incorrect—with epenthesis or deletion instead of a cluster. The presence of a vowel between two consonants was identified by the presence of a voice bar, clear first and second formants, and higher intensity relative to the flanking consonants. Speakers inserted a vocalic element between the consonants of a cluster 36% of the time. For instance, the English word 'stop' was often pronounced as shown in Figure 3.2, with an acoustic vocalic element 'V' inserted between the consonants [s] and [t], as well as word-finally after [p] (stop→[sVt upV]). The consonants of the clusters were changed 10% of the time, for instance, 'please' was sometimes pronounced [pjes]. 24% of the time, the stimuli were pronounced without inserted vowels or consonant deletion. Additionally, some of the words, for example, 'club', were unfamiliar to the speakers: during the production task, they reported that they did not know these words, and therefore decided not to pronounce them when invited to do so. This was the case for 29% of the total number of words tested.

Figure 3.3 gives an overview of the production of those ten speakers who reported low English proficiency, showing the frequency of different outputs across all stimuli for each speaker.

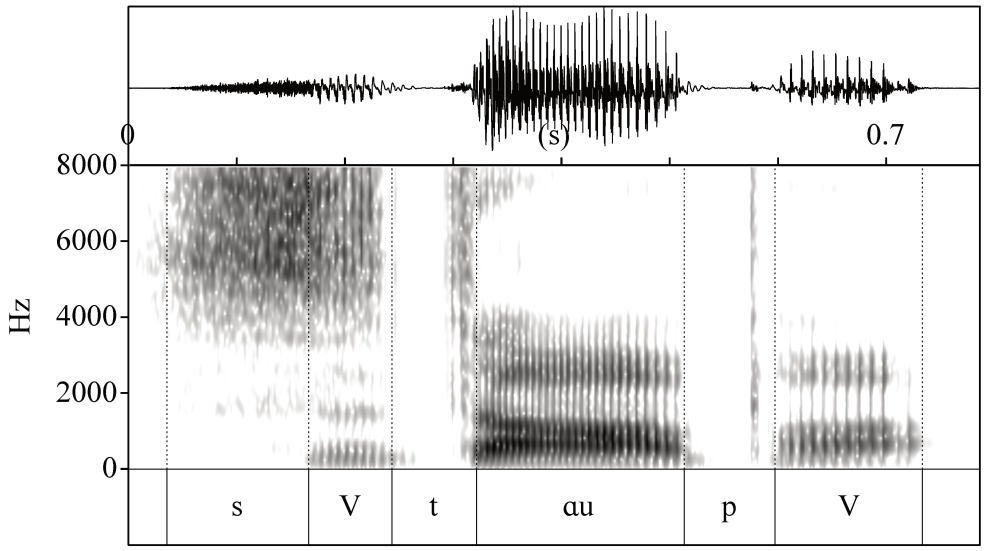


Figure 3.2: Waveform and spectrogram of the English word 'stop' produced by a monolingual Mandarin speaker, illustrating the occurrence of vocalic elements, labeled [V], between two consonants [s] and [t], and after [p].

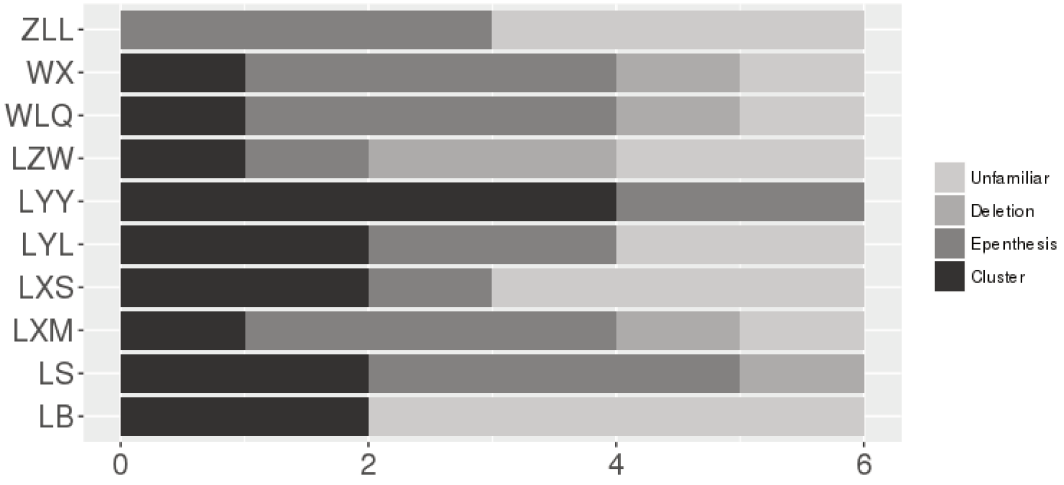


Figure 3.3: Frequency of different outputs across all English words for 10 participants with low English proficiency (results grouped by participants)

3.1.2 Stimuli

The oral English proficiency test was followed by a discrimination experiment, similar, but not identical to, the one presented in Chapter 2. Sixteen disyllabic target stimuli were constructed; the complete list of stimuli is shown in Table 3.1. All stimuli were Russian non-words, but they were judged phonotactically possible by a native speaker of Russian. In addition to the voiceless stop-stop (SS) clusters (e.g., /pt/) used in the ABX discrimination task, the stimuli included voiceless stop-nasal (SN) clusters (e.g./kn/), and voiceless stop-liquid (SL) clusters (e.g., /kl/). These clusters were tested in both word-initial and word-medial position. Also included were liquid-stop (LS) clusters (e.g., /lk/) and nasal- stop (NS) clusters (e.g., /nk/), which were tested only in word-medial position, since LS and NS combinations with voiceless stops are impossible in Russian word-initially. The tested stop consonants included the voiceless stops /p, t, k/. The liquid consonant was always /l/, and the nasal consonant /n/.

Recall that all of the tested clusters except NS (/nk/) are illicit in Mandarin. NS was therefore used for comparison with the non-native consonant clusters. Control stimuli were included for comparison with the target stimuli. The controls contained a CVC sequence (a vowel between the two consonants of the cluster).

The vowels surrounding C1C2 in VC1C2V sequences—/a/, /i/, and /u/—were chosen by the Russian speaker who produced the stimuli because they were the most natural for her to produce in the context of each cluster. This was done to ensure that the Russian pseudo-words were used sounded as native as possible, and that the speaker was comfortable pronouncing all the stimuli. With regard to word-level prosody, word-initial CC sequences always preceded the main stress, while those in medial position always followed the main stress. The vowel in the CVC control stimuli was always /a/. Recall that the vowel /a/ between two

Table 3.1: List of stimuli (/á/ denotes a stressed vowel)

	C1/C2 manner	Non-native clusters		Native cluster
		Word-initial/pre-stress	Word-medial/post-stress	Word-medial
Target stimuli (CC)	SS	tkápa, ktápa, tpáka	átka, ákta, átpa, ípta	únka
	SN	knápa	áknu	
	SL	klápa pláka	áklu ípla	
	LS		álka álpa	
Control stimuli (CVC)	SS	tɐkápa, kɐtápa, tɐpáka	átəka, ákəta, átəpa, ípəta	únəka
	SN	kɐnápa	ákənu	
	SL	kɐlápa pɐláka	ákəlu ípəla	
	LS		áləka áləpa	

consonants in the control items was never stressed, and was therefore produced as a reduced, centralized vowel (Hamilton, 1980). In pre-stressed syllables, the surface realization of unstressed/a/ is reported to be a low mid central [ɐ], and in post-stressed and unstressed syllables, it is a mid central [ə] (Padgett and Tabain, 2005).

All stimuli were recorded in a sound-treated room at Université Paris Diderot by a 23-year-old female native speaker of Russian. She had been living in France for two years at the time of the recording. The speaker reported that she was highly proficient in French, but continued to speak Russian regularly (at least two hours per day). The stimuli were produced in carrier sentences. For word-initial cluster sequences (e.g., /ktapa/), the carrier sentence used was as follows (where _ indicates the position of the non-word): Мой _ плывёт по морю [moj _ plʲɨvʲˈɵt pə ˈmorʲu] 'My_is swimming in the sea'; For stimuli with word-medial cluster

sequences (e.g., /akta/), the carrier sentence was used: **Наш _ плывёт по морю** [naʂ _ plɨvʲˈɵt pə ˈmorʲu] 'Our _ is swimming in the sea'. The word [naʂ] was preferred in word-initial cluster sequences because all forms containing word-medial clusters in the stimuli are vowel-initial (Table 3.1). The presence of a C-V boundary is meant to facilitate acoustic segmentation. Likewise, [moj] was preferred before the C-initial forms containing clusters. All sentences were written in Cyrillic and presented on a computer screen one at a time using PowerPoint. Each single sentence appeared randomly three times in total over the entire recording session. The Russian speaker was asked to read each sentence one by one naturally. She was instructed to make sure she pronounced the non-words exactly as if they were real Russian words. The stimuli were digitally recorded onto a laptop computer using the Praat program (Boersma and Weenink, 2014) and an AT2020 microphone, with a sampling rate of 44.1 kHz. Two (out of three) repetitions without production errors or peculiar intonation patterns were selected to be used as experimental stimuli. All stimuli were normalized using the 'scale intensity' function in Praat to obtain a mean intensity of 70 dB. A phonetician and native speaker of Russian listened to the recordings and judged whether each stimulus had been correctly pronounced. All tokens were approved.

3.1.3 Procedure

Participants were tested individually in the same sound-attenuated booth at the Phonetics and Speech Science Laboratory of the Chinese Academy of Social Sciences in Beijing. Prior to the experiment, they were told that they would be hearing words in a foreign language. Each stimulus was presented twice, with an inter-stimulus interval (ISI) of one second. After hearing each stimulus twice, the participants were asked to write it down on an answer sheet, using both Pinyin and a free style of transcription. Eight of the 21 participants, who knew how to

transcribe tones in Pinyin, also transcribed the tones they heard.

The complete instructions, which were given in Mandarin, are shown below with the English translation:

> “在接下来的实验中，您将听到一些外语单词。请用拼音写下您所听到的单词。这个实验不是为了测试您的外语水平。点击屏幕上的 “下一个词点击这里” 来继续。”
>
> “In the following task, you will hear words in a foreign language. Please transcribe what you hear in Pinyin. This experiment is not intended to test your linguistic skills. Please respond as quickly and accurately as possible. Click the button on the screen labeled “click here for the next item” to continue.”

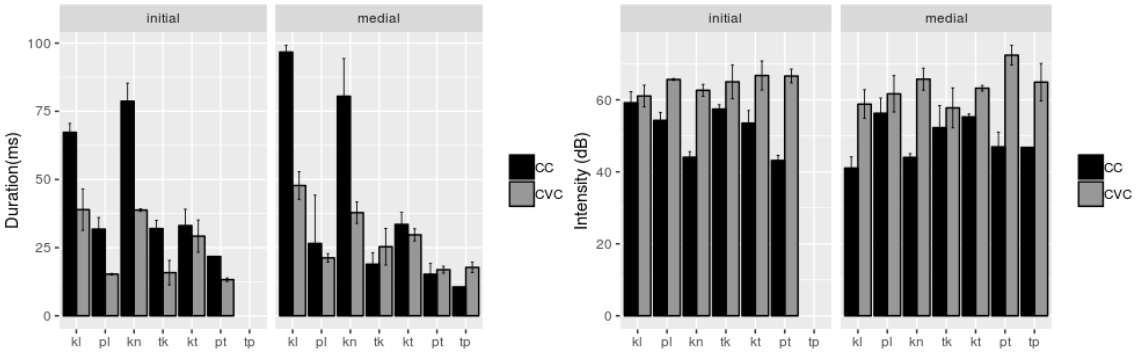
The stimuli were played through headphones. They were presented using the Multiple Forced Choice (MFC) function in Praat. The task was self-paced: after transcribing one item on the answer sheet, participants could proceed to the next stimulus by clicking the button on the screen labeled “click here for the next item”. Participants were first familiarized with the task through ten practice trials consisting of randomly-chosen items from the main experiment. No feedback was provided during the practice session. The experimental trials directly followed the practice trials, without any interruption. The test stimuli were repeated twice, each time in different random orders without any break between repetitions. The total number of stimuli presented to participants was 128 ((16 target sequences + 16 controls) $*$ 2 tokens $*$ 2 repetitions).

3.1.4 Acoustic analyses of the Russian stimuli

When analysing the results of the ABX experiment, we measured C1 burst duration and intensity, as shown in Figure 3.4, to determine whether speakers were

sensitive to these acoustic details in the transcription task. All measurements were made in Praat. The dynamic range for spectrogram display was set to the default value of 70 dB.

Figure 3.4: Duration (a) and intensity (b) of C1 burst in Russian stimuli for non-native CC clusters and CVC controls per cluster type in word-initial and word-medial positions. Error bars represent standard deviations.



(a) Duration of C1 burst

(b) Intensity of C1 burst

3.1.5 Coding for transcription

Responses were coded for the type of modifications made to the CC sequences in the transcription, according to the coding system summarized in Table 3.2. The same coding system was applied to both target CC and CVC control items. With regard to each strategy, a value of 1 in a column indicates that this strategy is present, and 0 indicates it is absent. In the case of multiple modifications per token, each one was labeled separately. For instance, both epenthesis and C1 change were coded when the token /ktapa/ was transcribed as <sitapu>.

Table 3.2: Coding for transcription

Coding	Criteria	Examples
Without V (correct)	C1C2 transcribed correctly, without any modifications.	/ktapa/ → <ktapa>
With vowel	A vowel was transcribed between target C1C2 or control C1VC2.	/ktapa/ → <ketapa>
C1 deletion	First consonant of C1C2 deleted.	/ktapa/ → <tapa>
C2 deletion	Second consonant of C1C2 deleted.	/ktapa/→ <kepa>
C1 change	First consonant of C1C2 has different place and manner, but same voicing.	/ktapa/ → <stapa>
C2 change	Second consonant of C1C2 has different place and manner, but same voicing.	/ktapa/ → <kpapa>

3.1.6 Statistics

We conducted Bayesian multinomial (polytomous) logistic regression models fitted to the transcriptions, since the coded modifications of the transcription are drawn from an unordered set of categories. The correct transcriptions (without V, e.g., /ktapa/ → <ktapa>) were excluded from the statistical analysis since they represented only 0.6% of the total number of transcriptions. Therefore, the dependent variables were multiple unordered modifications: With vowel, C1 deletion, C2 deletion, C1 change, C2 change. The reference modification selected as a baseline for comparison with other modifications was the insertion of a vowel in the cluster.

The independent variables were NATIVENESS (non-native CC clusters vs. native CVC sequences) and WORD POSITION (initial vs. medial). In addition, two phonetic features of the Russian stimuli, C1 BURST DURATION and C1 BURST INTENSITY, were included as independent variables for clusters where C1 is a stop. For contrast coding of categorical variables, the fixed factor was coded using sum-to-zero contrast coding such that each factor had a mean of zero, by assigning '1' to one level of the factor and '−1' to the other level. Random intercepts and slopes corresponding to the fixed effects were included for participants and items as permitted by the experimental design.

Statistical analyses were performed using the MCMCglmm package (Hadfield et al., 2010), a package for fitting Generalised Linear Mixed Models using Markov chain Monte Carlo techniques in R. The Bayesian prior on coefficients and other settings of the models were used as default settings for multinomial generalized linear mixed models (see Hadfield et al., 2010). We reported a 95% highest posterior density (HPD) interval for each coefficient and related p-values. For further analysis, when necessary, binomial logistic regressions were also performed using MCMCglmm.

3.2 Results

3.2.1 General results

Figure 3.5 shows the results of the transcription of clusters by Mandarin speakers. As already explained, participants correctly transcribed the non-native consonant clusters, without any modifications, only 0.6% of the time. Modifications in the transcription of the non-native consonant clusters included inserting a vowel in the cluster (62.3% of the total responses), deleting C1 (12.8%), deleting C2 (11.5%), changing C1 (23.1%), and changing C2 (18.5%). CVC controls were correctly

transcribed with vowels 93.2% of the time, and the proportion of modifications was much lower: C2 change (16.4%), C1 change (6.8%), C2 deletion (4.6%).

The analysis included fixed effects of NATIVENESS (CC & CVC) and WORD POSITION (initial & medial), as well as random intercepts and slopes for participants (NATIVENESS and POSITION) and items (NATIVENESS and POSITION). The model shows that transcribing a vowel in the cluster was the most prevalent response. Participants more often transcribed clusters with a vowel between the two consonants than using any other modifications, and accordingly there were significant negative coefficients for all modifications other than vowel insertion (*C1 deletion*= -2.3, *C2 deletion* = -4.7, *C1 change* = -2.2, *C2 change* = -3.1, all $p's < 0.05$). There was a significant difference between the non-native CC and the CVC controls for C1 deletion, but not for the other modifications. The transcription of the CVC controls involved significantly less C1 deletion than the transcription of non-native CC clusters (*C1 deletion* × CVC = -6.3, *95% HPD* $[-9.82, -2.87], p = 0.01$). There was no significant effect of word position on the perception of clusters.

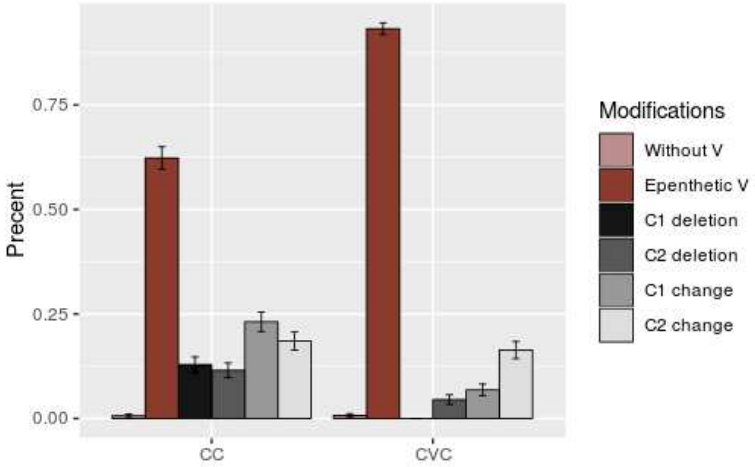


Figure 3.5: Proportion of outputs within non-native consonant clusters (CC) and controls (CVC). Error bars indicate 95% bootstrap confidence intervals.

We also found that the native cluster /nk/ was correctly transcribed 58.3% of the time, and only 4.8% of the time with a vowel. However, in the word /unka/, the alveolar /n/ was transcribed with a velar nasal /uŋka/ 32.1% of the time. This is consistent with Mandarin phonotactics, where a velar nasal precedes a velar stop.

We analyzed the stop-stop (SS), stop-nasal (SN), stop-liquid (SL), and liquid-stop (LS) clusters separately as shown below. We considered only clusters where C1 is a stop (SS, SN and SL) to determine the effect of C1 burst release on the perception of non-native CCs. Moreover, since only one type of SN cluster, /kn/, was used in the experiment, the analysis was restricted to SS and SL clusters. Given that participants in the discrimination task were able to discriminate between SS clusters and their CVC counterparts (e.g., /tkápa/ vs. /tɐkápa/), a separate analysis was carried out to understand whether this successful discrimination was due to the fact that participants perceptually modify the stop-stop clusters other than by inserting a vowel between the two consonants.

3.2.2 Stop-stop clusters

Figure 3.6 shows the frequency of modifications for different types of SS clusters. The analysis of SS clusters included fixed effects of NATIVENESS (CC & CVC), WORD POSITION (initial & medial), C1 BURST DURATION, and C1 BURST INTENSITY, as well as random intercepts and slopes for participants (NATIVENESS, WORD POSITION, C1 BURST DURATION, and C1 BURST INTENSITY) and items (NATIVENESS and WORD POSITION) (Items were nested within C1 BURST DURATION and INTENSITY, therefore only slopes for NATIVENESS and WORD POSITION were included). The model shows that C1 was more often deleted in non-native CC clusters than in CVC sequences (*C1 deletion* × CVC = −8.93, 95% HPD $[13.27, -4.14], p < 0.0001$) and more often

in word-initial than in word-medial position (*C1 deletion* × medial = −3.95, 95% HPD $[-6.15, -1.94], p = 0.0017$). This is because C1 in C1C2 clusters lacks formant transitions from the preceding segment, especially so in word-initial position.

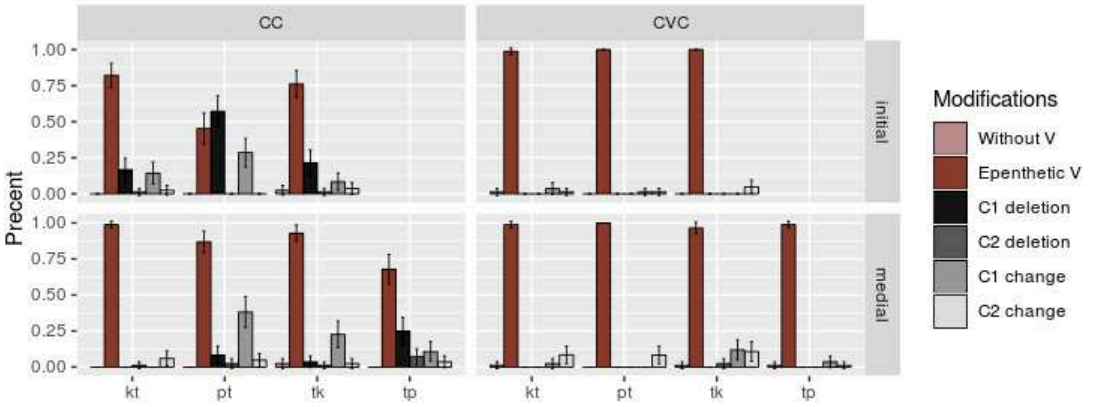


Figure 3.6: Proportion of outputs for each stop-stop cluster across word-initial and word-medial positions within CC and CVC conditions. Error bars indicate 95% bootstrap confidence intervals.

In addition, participants were sensitive to C1 burst intensity. The lower the burst intensity of C1, the more often C1 was deleted or changed. Conversely, the higher the burst intensity of C1, the more often CCs were transcribed with a vowel, as indicated by the significant negative coefficients comparing the transcription *with vowel* to other incorrect transcriptions (*C1 deletion* × *C1 intensity* = −0.2, 95% HPD $[-0.37, -0.02], p = 0.02$; *C1 change* × *C1 intensity* = −0.18, 95% HPD $[-0.29, -0.07], p = 0.001$). This explains why the rate of C1 deletion and C1 change varied across the place of articulation of C1. The labial consonant C1-*p* was more often deleted or changed than the velar C1-*k* and alveolar C1-*t*. Overall, C1 was omitted or changed more frequently in the word-initial cluster /pt/ (85.6% of the time) than in word-medial /pt/ (46.3%), followed by medial /tp/ (36%), initial /kt/ (31%), initial /tk/ (30%), medial /tk/ (26.1%). Word-medial /kt/ was always transcribed correctly. It is well known that stop consonants differ in

the energy of the released airflow depending on the place of articulation (Stevens, 1998). In the case of our study, for instance, the energy of the release for a voiceless bilabial plosive /p/ is weaker than it is for velar /k/ and alveolar /t/ (see 3.4). Our results show that C1 was never deleted for medial-/kt/, and that the labial stop /p/ in initial-/pt/ was the most frequently omitted C1. Failure to perceive /p/ can be attributed to its relatively weaker release burst. Moreover, when the stop is released into a constriction rather than a vowel, the release burst is the only information available as to the presence of the stop, as there are no formant transitions into a following vowel. When the cluster is word-initial, there are no preceding formant transitions either. All of these observations taken together may explain why the labial stop is the least perceptible C1 in a word-initial cluster.

In addition, the labial stop /p/ in /pt/ was transcribed with different consonants across word positions. The cluster /pt/ was transcribed as <sit> in word-initial position in 54% of all C1 change cases and as <fut> in word-medial position (62% of C1 change cases). C1-/p/ was always transcribed as a voiceless fricative, but with different place of articulation—either alveolar fricative /s/, or labiodental /f/. As mentioned above, since there were no transitions from the preceding formant for word-initial C1s, the C1 was released into a constriction of the following alveolar stop /t/. Therefore, the place of articulation of C1 was assimilated to the following alveolar stop, and transcribed as /s/. However, in word-medial position, the formant transitions of the preceding vowel and the labial /p/ led participants to perceive C1 as a labial consonant, in which case it was transcribed as labiodental /f/. As discussed above, CCs were always transcribed with a vowel between the two consonants C1C2. The vowel following alveolar /s/ was transcribed with the letter <i> in Pinyin, which corresponds to the dental approximant [ɹ̪] (Lee-Kim, 2014), while the vowel following /f/ was transcribed as <u>, which corresponds to the close back rounded vowel [u]. This shows that the vowels inserted in C1C2

clusters were affected by the place of articulation of the preceding consonant.

This finding is consistent with the results of the ABX discrimination experiment. Indeed, on the one hand, vowels were more often inserted in the transcription of /kt/ than of other clusters, suggesting it is especially hard for Mandarin speakers to discriminate between /kt/ and /kət/. On the other hand, the first consonant in the cluster /pt/ was more often deleted or changed than in other clusters, and the vowels perceived between /p/ and /t/ were different from a schwa, which suggests that Mandarin speakers had an easier time distinguishing between /pt/ and /pət/.

3.2.3 Stop-liquid clusters

Figure 3.7 shows the frequency of modifications for different types of SL clusters across different word positions in the two conditions CC and CVC.

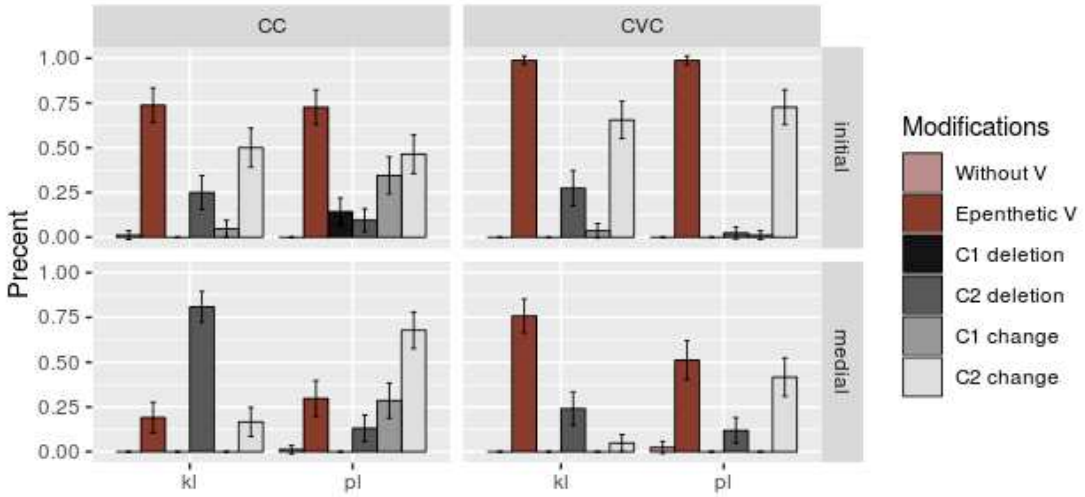


Figure 3.7: Proportion of outputs for each stop-liquid cluster across word-initial and word-medial positions in CC and CVC conditions. Error bars indicate 95% bootstrap confidence intervals.

The analysis of the SL clusters included fixed effects of NATIVENESS (CC

& CVC), WORD POSITION (initial & medial), C1 BURST DURATION, and C1 BURST INTENSITY, as well as random intercepts and slopes for participants (NATIVENESS, WORD POSITION, C1 BURST DURATION, and C1 BURST INTENSITY) and items (NATIVENESS and WORD POSITION) (Items were nested within C1 BURST DURATION and INTENSITY, therefore only slopes for NATIVENESS and WORD POSITION were included). The model shows C2 change was the most frequent modification. It occurred more often than the baseline, vowel insertion (*C2 change* = 9.9, *95% HPD* [1.79, 19.7], $p = 0.0212$). The liquid was often transcribed as a glide or a vowel, for instance, /klapa/ → <kewapu>; /ipla/ → <ipu>. Furthermore, the liquid of SL clusters was changed in both CC and CVC conditions to a labio-velar glide /w/ in word-initial position, and to the vowel <u> in word-medial position. Both transcriptions most likely stem from the fact that the liquid in Russian is a dark /ɬ/. Acoustically, dark /ɬ/ and /u/ are very similar, both sharing a close distance between F1 and F2 (Sproat and Fujimura, 1993). This acoustic similarity may be responsible for the frequent perception of /ɬ/ as a back rounded vowel or glide.

However, C1 burst duration and intensity did not affect the perception of non-native SL clusters. Lower amplitude of SL clusters did not result in more C1 deletion or C1 change as it did for SS clusters. This is probably due to the fact that C2, as a liquid consonant, involves a lower degree of constriction than a stop. The C1 stop is therefore not released into a full constriction, but into a more open vocal tract. It is thus less likely to be misperceived.

3.2.4 Stop-nasal clusters

Figure 3.8 shows the frequency of observed modifications for the SN cluster /kn/ across different word positions in clusters (CC) and controls (CVC).

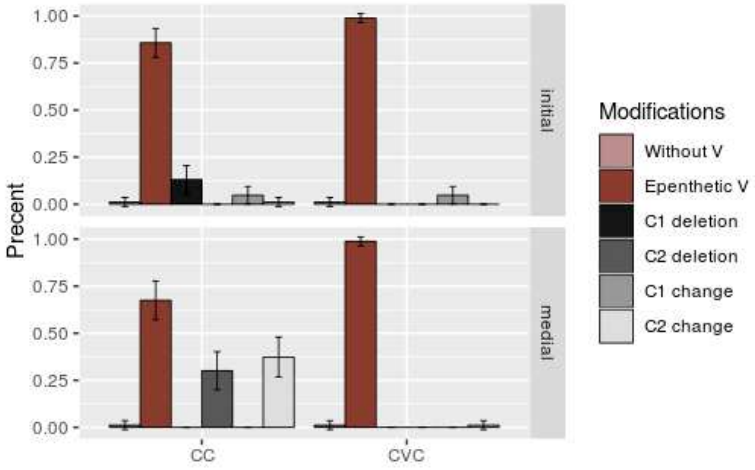


Figure 3.8: Proportion of outputs for the stop-nasal cluster /kn/ across word-initial and word-medial positions in CC and CVC conditions. Error bars indicate 95% bootstrap confidence intervals.

The analysis of the SN clusters included fixed effects of NATIVENESS (CC & CVC) and WORD POSITION (initial & medial), as well as random intercepts and slopes for participants and items (NATIVENESS and WORD POSITION). The model shows that there was a significant effect of NATIVENESS, where /kn/ had significantly more C2 deletion and C2 change than its control /kən/ (*C2 deletion* $\times$ CVC $= -8.9$, 95% HPD $[-15.72, -3.48], p = 0.002$; *C2 change* $\times$ CVC $= -6.3$, 95% HPD $[-11.51, -1.21], p = 0.01$).

For C1 deletion, C1 change, and C2 change in the transcription of non-native CCs, modification type and frequency of occurrence varied depending on whether [kn] was initial or medial. Initial-/kn/ was modified by C1 deletion 13% of the time. Medial-/kn/ was modified by C2 deletion 30% of the time, and by C2 change 37% of the time. Mixed-effects binary logistic regressions of different modifications against cluster type (word position) were performed to better understand this difference.

We found that vowels were more frequently inserted in initial-/kn/ clusters

than in medial-/kn/ clusters, as indicated by the significant negative coefficients for medial-/kn/ (medial-/kn/ = -4.5, $p's < 0.05$). In addition, C1 deletion was more frequent for the initial-/kn/ than for medial-/kn/ (initial-/kn/ = 28.7, $p < 0.001$). C2 deletion was less frequent for initial-/kn/ than for medial-/kn/(initial-/kn/ = -38, $p's < 0.001$). C2 change was also less frequent for initial-/kn/ than for medial-/kn/ (initial-/kn/ = -9.3, $p's < 0.001$).

3.2.5 Liquid-stop clusters

Figure 3.9 shows the frequency of the modifications for different types of LS clusters in word-medial position in CC clusters and CVC controls. We observed that unlike with other non-native clusters, vowels were not the most common transcription for LS. Instead, the most frequent modification was C1 change.

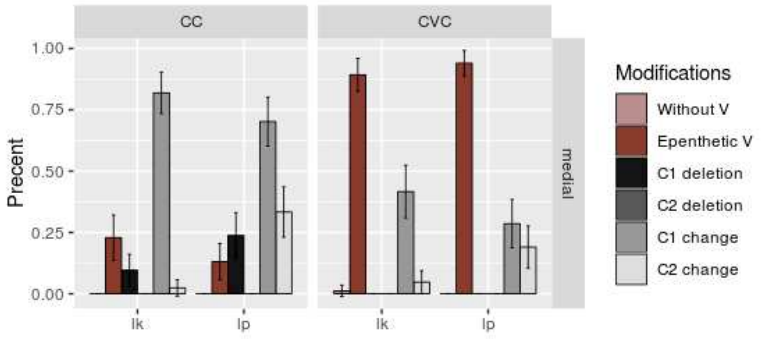


Figure 3.9: Proportion of outputs for each liquid-stop cluster in word-initial and word-medial position within CC and CVC conditions. Error bars indicate 95% bootstrap confidence intervals.

The analysis of the SN clusters included fixed effects of NATIVENESS (CC & CVC) and WORD POSITION (initial & medial), as well as random intercepts and slopes for participants and items (NATIVENESS and WORD POSITION). Since there was no C2 deletion at all, we removed this modification from the dependent variable. Recall that the reference level was vowel insertion. Results

showed that C1 change was the most frequent transcription, as indicated by a significant positive coefficient for C1 change ($C1\ change = 1.7$, $p < 4e-04$). For instance, /alpa/ was more often transcribed <aupa>.

The liquid in LS combinations (e.g., /alәpa/) is a dark liquid [ɬ] in Russian. As in the case of SL clusters, the dark [ɬ] was most often transcribed as the vowel /u/ in word-medial position.

3.2.6 Transcription choices for vowels

Figure 3.10 shows the vowels used to transcribe the vowels in non-native CC clusters, plotted by the first consonant (C1).

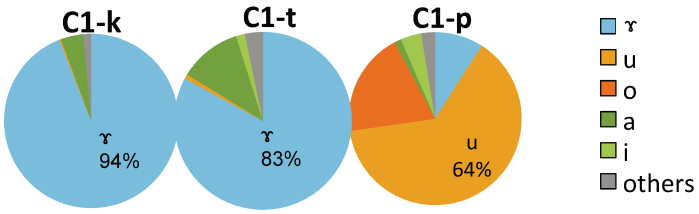


Figure 3.10: Vowels used to transcribe the vowels plotted by first consonant.

A multinomial (polytomous) logistic regression model was fitted to the transcribed vowels with C1 IDENTITY as fixed factor, as well as random intercepts and slopes corresponding to the fixed effect for participants and items. The model reveals a significant effect of C1 IDENTITY on the transcribed vowels. A back mid unrounded vowel [ɤ] (<e> in Pinyin) was the most frequently-used vowel to transcribe the vowels inserted after velar and alveolar stops. In Mandarin, [ɤ] is an allophone of the mid vowel /ə/, which is the most neutral vowel of Mandarin. The mid vowel can be deleted in fast conversational speech when it precedes a nasal consonant (Weinberger, 1997). In the perception of non-native consonant

clusters, the phonetic cue between two consonants in CCs was mapped onto the acoustically closest vowel, [ɤ], in Mandarin Chinese.

However, a different vowel was used after the labial stop /p/. For C1-/p/, the most frequently-used vowel was the back rounded vowel [u], transcribed <u> ($C1-p\times[u] = 15.3$, 95% HPD [6.7, 23.5], $p = 3e-04$). It should be noted that this may be due to the phonological restrictions on syllable structure in Mandarin, which disallow *[pɤ], but allow [tɤ] and [kɤ] (see Lin, 2007). Moreover, the sequence <pe> is never used in Pinyin transcription, leaving only the following options for transcribing a vowel after /p/: [pi, pa, pu]. The final choice was a close back rounded vowel [u], which agrees with the preceding labial consonant in place of articulation. Since the vowel shares the same place of articulation as the consonant, it is minimally obtrusive and helps make the Mandarin output (/put/) as similar as possible to the input (/pt/) (see also Miao, 2005).

This finding may further explain why Mandarin listeners were highly sensitive to the difference between /pt/ and /pət/ in the ABX discrimination task. Since the vowel in the controls was a schwa, listeners were sensitive to the difference between the vowel perceived inside the cluster /pt/, which was a back close rounded /u/, and the vowel in the control sequence, which was a schwa /ə/.

3.3 Interim summary

In the transcription experiment, we observed that monolingual speakers of Mandarin Chinese always transcribed non-native CC clusters with a vowel between the two consonants. However, they were also sensitive to the intensity of the first consonant (C1) burst in the Russian stimuli, since lower burst amplitude of C1 made participants less likely to modify the initial stop by inserting a vowel before C2. When C1 intensity was lower, C1 was more often deleted or changed. In

addition, a mid back unrounded vowel [ɤ] (spelled <e> in Pinyin) was the most frequently used vowel to transcribe the vowels, except for C1-/p/. The choice of the vowel [ɤ] is due to the fact that it is the most neutral and centralized vowel in Mandarin (see also Durvasula et al., 2018; Miao, 2005), and therefore the closest vowel to the transition between C1 and C2 in a cluster. However, due to phonological restrictions on syllable structure in Mandarin, which disallow *[pɤ], the vowel occurring after C1-/p/ was a close back rounded /u/, which agrees with the place of articulation of the previous labial stop. These findings are consistent with the results of the ABX discrimination experiment, which showed that Mandarin listeners were highly sensitive to the contrast between /pt/ and /pət/, since the cluster /pt/ was variously perceived with C1 deletion, C1 change, and a vowel, specifically a close back vowel /u/ instead of a schwa. Based on these findings, we conclude that phonetic details primarily contribute to the perception of non-native consonant clusters: monolingual speakers of Mandarin Chinese try to decode phonetic details of the input to map the sounds and structures as faithfully as possible to those of their native language in their output.

It is, however, very important to consider these results with great caution. We saw that Mandarin speakers are quite successful in discriminating between CC and CVC, which suggests that they do not systematically perceive a vowel in a non-native cluster. On the other hand, the transcription experiment revealed much higher rates of vowel transcription in clusters. While the transcription task served its purpose in revealing different types of modifications in the perception of clusters, it has very likely over-generated vowel transcription because of the rules of Pinyin orthography. In Pinyin, a vowel between two consonants is obligatory. Speakers may therefore have tried to relate their perception to possible spelling forms in the Pinyin system. Thus, the transcription task alone does not allow us to reliably test the effect of phonetic details on the insertion of vowels between two

consonants. For this purpose, we used a repetition experiment, which we present in the next section. We will focus on the quality of the inserted vowels, and on the gestural coordination patterns involved in the production of non-native consonant clusters, as inferred from acoustic measures of overlap.

Chapter 4

Production of non-native CCs

In the previous perception experiments—ABX discrimination and transcription—the perception of non-native consonant clusters by monolingual speakers of Mandarin was found to be affected by both phonotactic knowledge and phonetic details of the inputs. The ABX discrimination task shows that Mandarin speakers had difficulty perceiving the difference between CC and CVC no more than 35% of the time. Their sensitivity to the CC and CVC contrast was significantly above chance. However, successful discrimination between CC and CVC does not necessarily show that Mandarin speakers did not perceive a vowel inside of non-native clusters and perceived non-native clusters 'correctly' . Results from the transcription task indicate that Mandarin speakers modified non-native consonant clusters in various ways, by inserting a vowel, deleting one of the consonants or changing a feature, depending on phonetic details in the stimuli. We saw that lower amplitude of C1 release results in more deletion of the first consonant, but higher amplitude of C1 release leads more often to the transcription of a vowel between two consonants. Thus, Mandarin speakers are sensitive to phonetic details, and try to decode phonetic details in the inputs to map the sounds and structures as faithfully as possible onto native ones. One question arises: what role does this

phonetic decoding process play in the production of non-native consonant clusters?

We hypothesized that if sensitivity to phonetic details primarily contributes to non-native production, then different types of modifications will be produced for different types of clusters.

Regarding the nature of the vowels occurring between two consonants, as discussed earlier, Davidson (2006a) claimed that it may not be a lexical vowel, but a 'transition' emerging from reduced overlap of the flanking consonants. We therefore need a neutral term to refer to such vowels. We will refer to them as 'vocoids' in the production.

Davidson found that, since a 'transition' lacks an articulatory target of its own, both the F1 and F2 values of the 'transition' were lower than those of a lexical vowel. It is a transitional movement from the preceding consonant to the following one such that the duration of the 'transition' is considerably shorter than it is for a lexical vowel. On the other hand, it is possible that speaker-listeners try to map the sounds in the inputs as faithfully as possible onto those of their native language. In the case of Mandarin Chinese, the native Mandarin sequence closest to a CC cluster would be one containing a reduced or a deleted vowel. As reviewed in Section 1.4, Lin and Yan (1980) found that a reduced vowel is 50% shorter than a full lexical vowel, and its quality is that of a centralized, schwa-like sound, that is, both F1 and F2 centralize to the middle of the vowel space.

We tested this hypothesis in a prompted production task. We analyzed the frequency of inserted vocoids and other modifications of CC clusters, the quality of inserted vocoids (duration and formants), and the temporal coordination of the flanking consonants based on acoustic measures of overlap. The production of non-native CC clusters was compared with the production of native CVC sequences. The experiment was run in Beijing, with the same participants and the same stimuli as the transcription experiment presented in Chapter 3. The

Mandarin speakers were asked to repeat the stimuli they heard. The prompted production task was conducted before the transcription experiment to ensure that the participants did not have any previous exposure to the non-native clusters. Without orthographic information and without training, speakers are more likely to rely primarily on phonetic decoding to produce non-native CCs.

4.1 Methodology

4.1.1 Participants

The participants were the same 21 monolingual speakers of Mandarin as in Experiment 2–transcription experiment (see Section 3.1.1).

4.1.2 Stimuli

The stimuli were the same as in the transcription experiment (see Section 3.1.2): 16 target CC clusters and 16 CVC controls.

4.1.3 Procedure

Participants were seated in a sound-treated booth in front of a computer running a Praat script. No orthographic information was displayed on the screen. The stimuli were played via headphones with an inter-stimulus interval of 1000ms. Participants heard each stimulus twice, then repeated it three times in a row. Participants were not allowed to change their production responses.

For the first repetition task, the instructions were given as follows in Mandarin (the English translation is provided below):

> “您将会听到一些外语单词，请仔细听到最后。每个单词播放两遍，然后请跟读三遍。这个实验不是为了测试您的外语水平。”

> *"In the following task, you will hear words in a foreign language. Please listen carefully until the end of the word. Each word will be played twice. Please repeat the words three times after hearing them. This experiment is not intended to test your language skills."*

The repetition list was played twice in different random orders. The total number of productions was 32 items $*$ 2 orders $*$ 3 repetitions = 192. The repetition task lasted about 35 minutes and was divided into 3 blocks. Participants were given 5-minute breaks between blocks. All repetitions were captured through an AKG C 4000b microphone and recorded using Praat onto a Soundcraft Spirit Studio console connected to a Windows computer. The WAV files were recorded at a 44.1 kHz sampling rate (16-bit resolution; 1-channel). 10 practice trials were given at the beginning in order to familiarize participants with the experimental setting.

4.2 Data analysis

1344 tokens were analyzed by listening to them and examining their waveforms and spectrograms in Praat. Each token consisted only of the first repetition produced by the participants in a series of three repetitions. The first repetition was chosen so that any potential effect of learning would be avoided. Acoustic details of the inserted vocoids, including duration and the first and the second formant (F1 and F2), were measured to understand the nature of the vocoids. The acoustic timing lag of the flanking consonants in the production of non-native consonant clusters was also measured. All measurements of the target CCs were compared with those of CVC control items. Moreover, each token was coded based on the type of modification that occurred, if any (insertion, consonant change, or consonant deletion). The measurements and coding were carried out by the author, who is a native speaker of Mandarin Chinese, and by one phonetically-trained research

assistant.

4.2.1 Criteria for classifying the productions

Figure 4.1 shows the spectrogram and waveform of three representative tokens in the production of non-native consonant clusters: acluster with an vocoid (Figure 4.1a), cluster without an vocoid (Figure 4.1b), and cluster with voicing between two consonants (Figure 4.1c).

4.2.1.1 With vocoid

Figure 4.1a shows a voiced vocoid inserted between two consonants C1C2, which, following Davidson et al. (2015); Wilson et al. (2014), is identified by:

- Voice bar
- Clear formant structure
- Higher intensity than the flanking sounds

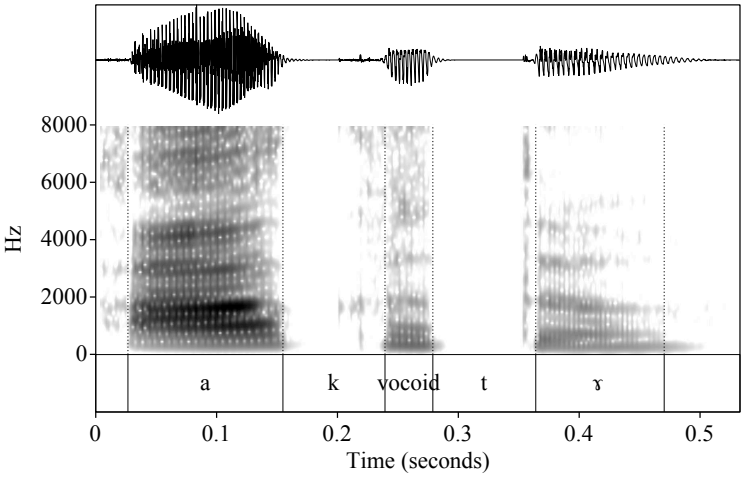
4.2.1.2 Without vocoid

Figure 4.1b shows a C1C2 sequence without no vocoid inserted. The absence of a vocoid is signaled by:

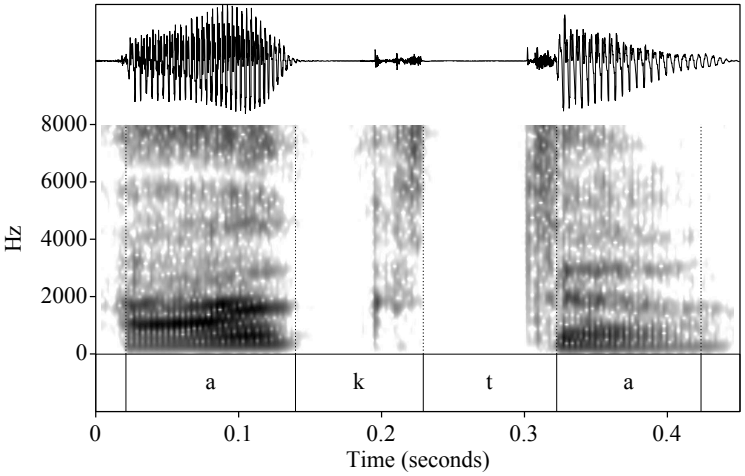
- No voice bar
- No formant structure

4.2.1.3 With voicing

Figure 4.1c shows a C1C2 sequence with voicing inserted between the two consonants, identified by:



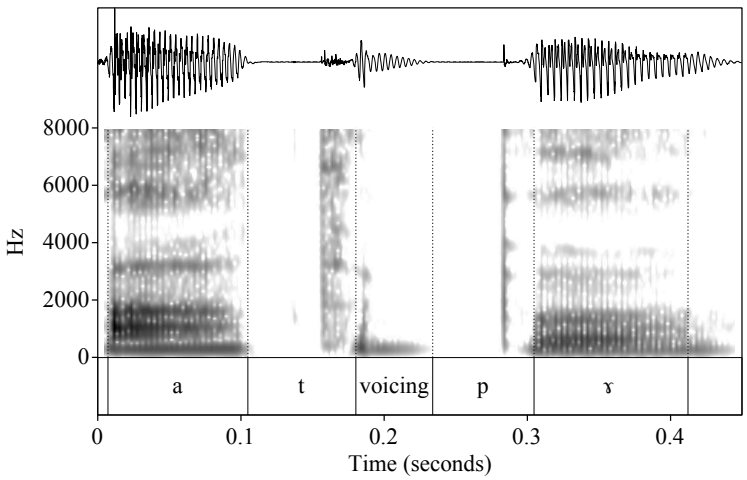
With vocoid



Without vocoid

- Voice bar

- No formant structure



With voicing

Figure 4.1: Spectrogram and waveform illustrating different productions of three sample tokens: with an vocoid ([akˀtɤ], top), without an vocoid ([aktɤ], middle), and with voicing ([at_pɤ], bottom).

4.2.2 Measurements for inserted vocoids

Duration For the purposes of vocoid segmentation, the onset of the vocoid was identified as the first zero-crossing point after the preceding consonant offset on the waveform. The offset of the vocoid is marked at the end of the visible second formant on the spectrogram, excluding the portion of the voice decay time (C→D, Figure 4.1). Harmonics-to-noise ratio (HNR) measurements were used to identify voicing in the signal (Krom, 1993). If there was no voicing in the signal, the HNR values could not be determined. The algorithm for calculating the HNR was based on a forward cross-correlation analysis, which was computed using Praat through the Harmonicity (cc) function. It was run with the following settings: time steps of 10ms, pitch floor of 75 Hz, pitch ceiling of 500 Hz, silence threshold of 0.1, and 1 period per window. The dynamic range for spectrogram display in Praat was set at the default value of 70 dB.

First and second formants (vocoid quality) The first two formants were extracted at the midpoint of the vocalic intervals using LPC (Linear Predictive Coding) analysis. The formant tracts were used with the default values in Praat, based on the Burg algorithm (Childers, 1978; Press et al., 1992), with a Gaussian window of 25ms, a time step of 10ms, and a pre-emphasis of 50 Hz. Formant tracks were laid over a broadband spectrogram with the following settings: Fourier method (FFT), Gaussian window of 5ms, time step of 2ms, maximum formant frequency of 5500 Hz for female speakers and 5000 Hz for male speakers, bandwidth of 250 Hz, and pre-emphasis of 6 dB/octave. In addition, since the inserted vocoids were sometimes very short (the shortest being 20ms long), the automatically extracted formants could be less reliable. Thus, the extracted formant values were manually checked by visually comparing formants from linear interpolation Burg LPC analysis and formants from spectrograms. When necessary, the parameters were modified to get an optimal fit. This was done using the Praat script cp_formants, written by Emmanuel Ferragne (available from https://tinyurl.com/hwv6a96). Since the current study involves both production and perception of non-native consonant clusters, taking into account the human auditory system, the formant values were converted from Hertz to a perceptual Bark Scale (Traunmüller, 1990).

4.2.3 Measurements for acoustic timing lags

Table 4.1 shows the three measurements used for acoustic timing lags in the production of non-native C1C2 consonant clusters, including the entire duration of C1C2, the inter-plateau interval (IPI) and the inter-burst interval (IBI). For relevant labels, see Figure 4.2.

The entire temporal interval of C1C2 was measured from the offset of the vowel

Table 4.1: Measurement criteria of acoustic timing lags

Measurements	Criteria	Labels
Entire C1C2	from the offset of the preceding vowel to the onset of the following (only for word-medial clusters)	A→F
IBI	from the onset of the release burst of C1 to the onset of the release burst of C2	B→E
IPI	from the onset of the release burst of C1 to target achievement of C2	B→D

preceding C1 to the onset of the vowel following C2 (cf. Ridouane and Fougeron (2011), for Tashlhiyt). The offset of the preceding vowel was identified as the end of the visible second formant on the spectrogram. This measurement was only applied to word-medial clusters, not word-initial ones, since there was no preceding vowel in that case. Thus, we only compared C1C2 duration for word-medial clusters. The onset of the following vowel was identified as the first zero-crossing point after C2 offset on the waveform. It should be noted that the following vowel was sometimes devoiced. For this reason, the onset of the following vowel was defined as being immediately after the burst (or sometimes multiple bursts) of C2.

Two acoustic measures of timing lag have been used in the literature: IPI, the inter-plateau interval of C1C2, and IBI, the inter-burst interval. The inter-plateau interval (IPI) is the interval from the release of C1 to the target achievement of C2. The more closely C1 and C2 are timed together (with higher overlap), the shorter this interval is. When the degree of overlap is low, the IPI interval corresponds to the open transition between the two consonants of a cluster. In this case it serves as an indicator for how long the vocal tract remains open between two

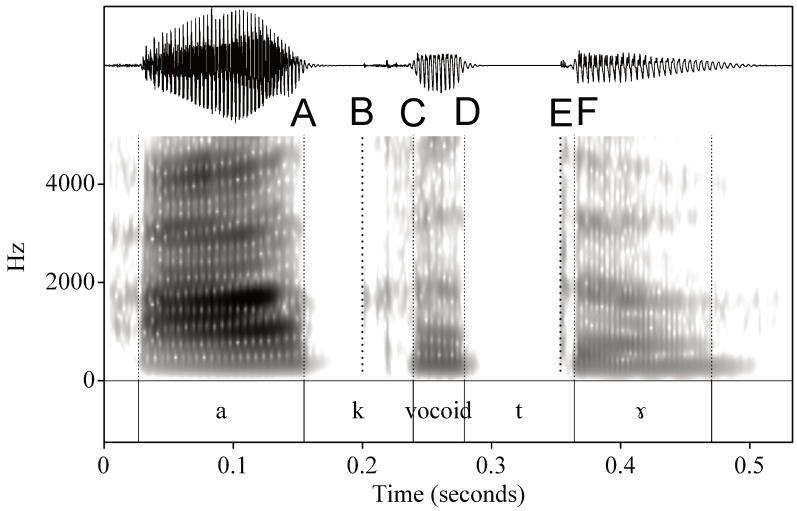


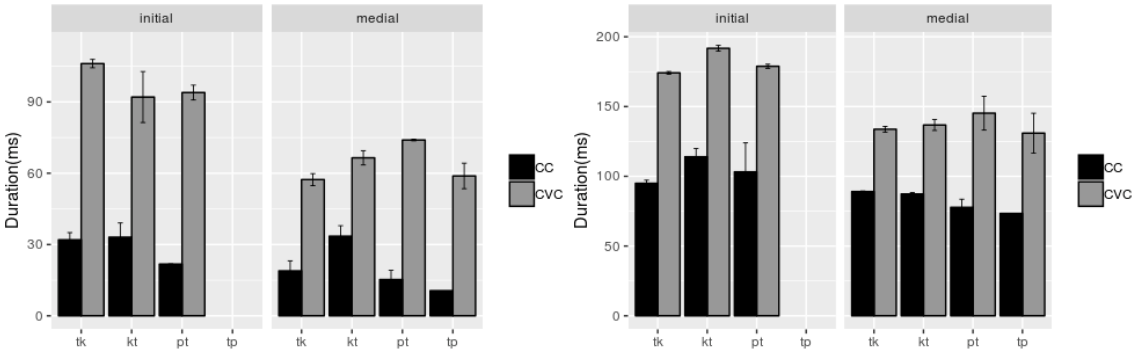
Figure 4.2: Measurements of inserted vocoid duration and CC acoustic timing lags. An example of the production of /akta/ by one participant, where an vocoid was inserted between /k/ and /t/. A: offset of preceding vowel, B: onset of C1 burst, C: onset of inserted vocoid, D: onset of C2 target achievement, E: onset of C2 burst, F: onset of following vowel

constrictions. During this open transition, a vocoid may be acoustically produced. The IPI of C1C2 was measured from the onset of the release burst of C1 to the onset of the target achievement (constriction) of C2 (see also Gafos, 2002; Shaw and Kawahara, 2018; Bellik, 2018). The left edge of the IPI interval was the onset of the release of C1 identified by an increase in amplitude, where C1 was audible and the release was visible in the waveform. The right edge of the IPI was either: 1) at the end of C1, when there was no vocoid inserted, identified by a decrease in amplitude until C1 was no longer audible or visible on the waveform; 2) at the end of the vocoid, identified by the end of the visible second formant on the spectrogram, when there was an vocoid inserted.

The inter-burst interval (IBI) of C1C2 was measured from the onset of the release burst of the first consonant (C1) to the onset of the release burst of the

second consonant (C2) (cf. Wright (1996) for Tsou, Chitoran (1999) for Georgian). It corresponds to the interval between the release of C1 and the release of C2. The onset of the release burst of C1 and C2 was identified by a high increase in amplitude, where the consonant was audible and the release was visible on the waveform. This is a different measure from the IPI, which does not give us any information about the presence or absence of an open transition. We consider it here as an acoustic measure of overlap that takes the constriction releases as reference landmarks, and indicates how far apart the two releases are. The longer the interval, the farther apart the releases, and the less overlapped the consonants are.

Acoustic timing lags were measured for the production of Mandarin speakers as well as for the Russian stimuli. Figure 4.3 shows the inter-plateau interval (IPI) and the inter-burst interval (IBI) in Russian stimuli for non-native CCs clusters and CVC controls per cluster type in word-initial and word-medial positions.



(a) IPI (b) IBI

Figure 4.3: (a) Inter-plateau interval (IPI) and (b) Inter-burst interval (IBI) in Russian stimuli for non-native CC clusters and CVC controls per cluster type in word-initial and word-medial positions. Error bars represent standard deviations.

4.2.4 Coding for productions

All productions of consonant clusters were coded based on the scheme laid out in Table 4.2. In the case of multiple modifications, each one was recorded in the coding, for instance, both epenthesis and C1 change were labeled in the token [ktapa] → [sɤtak^{h}]. Moreover, we also applied the same coding to the production of the control items (CVC). Unidentifiable modifications (less than 0.5% of tokens) were removed prior to analysis.

Table 4.2: Coding for productions

Coding	Criteria	Examples
Without V (correct)	C1C2 with no acoustic V inserted and no other modifications	[kt͡apa]→[kt͡apa]
With V	Vocalic element inserted between CC with clear formants, voice bar, and higher intensity than the flanking consonants	[kt͡apa]→[k^{hə}t͡apa]
With voicing	Voice bar, but no clear formants	[kt͡apa]→[k^{h}_t͡apa]
C1 Deletion	C1 deleted in C1C2	[kt͡apa]→[t͡apa]
C2 Deletion	C2 deleted in C1C2	[kt͡apa]→[k͡apa]
C1 change	Change in place or manner of articulation of C1 in C1C2	[kt͡apa]→[st͡apa]
C2 change	Change in place or manner of articulation of C2 in C1C2	[kt͡apa]→[kp͡apa]

4.3 Results

4.3.1 Frequency of inserted vocoids and other modifications

Figure 4.4 shows the frequency of the different outputs which occurred in the production of non-native CCs and the CVC controls after participants heard the Russian stimuli. Overall, Mandarin speakers modified the non-native consonant clusters by inserting an vocoid 26% of the time, and by inserting voicing between two consonants 7% of the time. Speakers also made other modifications, such as changing or deleting a consonant, about 56% of the time. Interestingly, unlike the results of the transcription task, participants produced the non-native consonant clusters correctly, without using any modifications, 13% of the time.

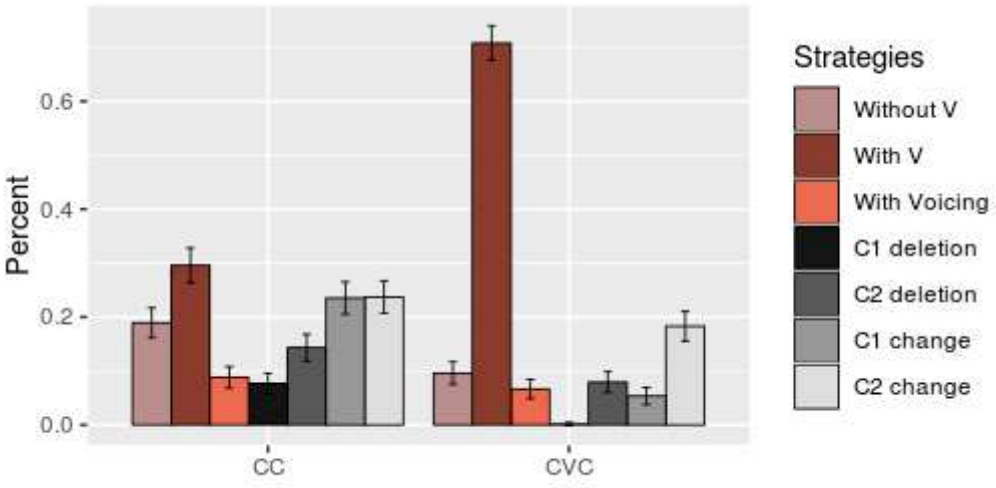


Figure 4.4: Proportion of outputs within both non-native consonant clusters (CC) and controls (CVC). Error bars indicate 95% bootstrap confidence intervals.

However, for the consonant cluster /nk/, the only consonant sequence that exists in Mandarin, participants inserted vocoids between the two consonants less

often, only 2% of the time. The sequence /nk/ is licit in word medial position, for example in [jæn.k^{h}ɤ] 'banquet'. Otherwise, as we observed in the transcription task, the most frequent strategy used for /nk/ was C1 change. The apical alveolar nasal [n] in the cluster [nk] was produced as a velar nasal [ŋ] 69% of the time.

Table 4.3 shows that non-native consonant clusters were modified with a variety of manner combinations. For instance, we observed that SS clusters were more often modified with vocoid insertion and C1 change, and SN clusters were more often modified with vocoid or without any modifications (correct production). SL clusters were more often modified with C2 change. We observed that in word-initial position, the liquid was commonly produced as a labio-velar glide [w], while in word-medial position, it was produced as a back closed vowel [u] or [o]. Similarly, in LS clusters, the liquid was always changed into a back vowel [u] or [o] 63% of time, and was deleted 30% of the time, with vocoids produced between the two consonants only 1% of the time. LS clusters were excluded from further comparisons.

The frequency information in Table 4.3 gives us an idea of vocoid insertion in the adaptation of non-native CCs, and of the other modifications which occur in the process. To understand how phonological knowledge and phonetic details influence vocoid insertion, we performed Bayesian multinomial logistic mixed-effects models using the MCMCglmm package in R. The dependent variables were multiple unordered modifications: *without V, with V, with voicing, C1 deletion, C2 deletion, C1 change, C2 change*. The reference strategy used as a baseline for comparison with other modifications was the production of a cluster without an inserted vocoid or any other modifications (*without V*). The independent variables were NATIVENESS (non-native CC vs. native CVC), WORD POSITION (initial vs. medial), C1/C2 MANNER (SL, SN, SS) and two factors pertaining to phonetic details in the stimuli: C1 BURST DURATION and C1 BURST INTENSITY.

Outputs	SL		SN		SS		LS		*NS*	
	CC	CVC	CC	CVC	CC	CVC	CC	CVC	CC	CVC
Without V (Correct)	13%	8%	29%	13%	20%	9%	1%	11%	23%	0
With V	26%	59%	34%	68%	30%	78%	1%	35%	2%	100%
With voicing	2%	4%	10%	11%	9%	7%	2%	1%	0	0
C1 deletion	0	0	6%	0	9%	0.3%	31%	8%	4%	0
C2 deletion	32%	25%	19%	1%	3%	0	1%	1%	0	0
C1 change	21%	2%	1%	2%	30%	8%	63%	48%	69%	0
C2 change	49%	47%	10%	8%	13%	4%	14%	7%	0	0

Table 4.3: Percentage of outputs in the production of CC clusters and CVC controls for each C1/C2 manner combination, where SL, SN, SS and LS are non-native clusters, and NS is a native cluster.

NATIVENESS and WORD POSITION are binary factors, C1/C2 MANNER is a three-level factor, and C1 BURST DURATION and C1 BURST INTENSITY are continuous factors. For contrast coding of categorical variables, all binary fixed factors were coded using sum-to-zero contrast coding such that each factor had a mean of zero, by assigning '1' to one level of the factor and '-1' to the other level. The 3-level factor, C1/C2 MANNER, was coded such that each level of the factor (SN, SS) was compared to the reference level (SL). The random effect included intercepts for participants and items.

Figure 4.4 shows the details of the different modifications that occurred for both non-native consonant clusters (CCs) and native controls (CVCs). The model with only Nativeness as the fixed factor shows that vocoids occurred significantly more often in the CVC controls (where they were present in the Russian stimulus) than in the non-native consonant clusters ($With\ V \times Nativeness = 0.36$, $95\%\ HPD$ $[0.07, 0.64]$, $p = 0.0147$). However, there were no significant differences between CC and CVC for the other modifications across different nativeness conditions.

In addition, the model with C1/C2 MANNER, WORD POSITION, C1 BURST DURATION and C1 BURST INTENSITY as fixed factors revealed a significant effect of C1/C2 MANNER, indicated by a significant interaction between C2 change and SS and SN ($C2\ change \times SS = -0.58$, $95\%\ HPD$ $[-0.82, -0.056]$, $p = 0.01462$; $C2\ change \times SN = -0.64$, $95\%\ HPD$ $[-1.26, -0.02]$, $p = 0.03931$, respectively). In SL clusters, since the liquid is a dark liquid [ɬ] in the Russian stimuli, the dark [ɬ] was changed into a labiovelar glide [w] when it was in word-initial position (klapa→kawapa/kwapa) or into a close back rounded vowel [u] or a mid-close back rounded vowel [o] in word-medial position (ipla→ipo). This is due to the fact that acoustically, F1 and F2 of both dark [ɬ] and the close back rounded vowel are very close (Sproat and Fujimura, 1993). This result also explains why fewer vocoids were inserted between the two consonants of SL clusters: they were perceived as

CG-initial clusters and as CV syllables, both of which are allowed in Mandarin. These productions are consistent with the results in the transcription task.

There was also a significant effect of C1 BURST INTENSITY. This is reflected in the model by a significant interaction between C1 change and C1 burst intensity (*C1 change* × *Burst* = -0.029, *95% HPD* $[-0.05, -0.006]$, p = 0.01448). For example, the most frequent C1 change in word initial position occurred with C1-/p/ clusters, for example [ptaka] → [stakɤ], where the average intensity of C1-/p/ was 43 dB ($SD = 1.4$). In word medial position, C1-/t/ clusters were changed more often, for example, [atka] → [afkɤ], where the average burst intensity of C1-/t/ was 52 dB ($SD = 6.1$). This result is consistent with the findings of Wilson et al. (2014).

Therefore, we conclude that the adaptation of non-native CCs by Mandarin speakers was influenced both by their knowledge of native phonotactics and by phonetic details of the stimuli. Most often they inserted a vocoid between two consonants, consistently with their native phonotactics. However, various modifications in their production indicate that they relied on the decoding of phonetic details to produce non-native CCs. Lower C1 burst amplitude resulted in more C1 change. Next, we focus on the productions involving inserted vocoids, and we examine the quality of the vocoids.

4.3.2 Acoustic properties of inserted vocoids

To find out more about the nature of the vocoids inserted in the adaptation of non-native CCs, we analyzed their acoustic properties, including duration and vowel space (first and second formants, henceforth F1 and F2). We compared the quality of the vocoids Mandarin speakers produced between two consonants when prompted by non-native CCs, with that of the vowels produced when prompted by control CVC sequences.

For the comparison of duration, linear mixed-effects models were computed using the *lme4* package in R. The test independent variable was NATIVENESS (CC vs. CVC), which was intended to compare the inserted vocoids in CCs to the vowels in CVCs. We included a control variable WORD POSITION (initial vs. medial). This is based on the observation that in the Russian recordings, the vowel duration in CVC sequences varies across word positions, that is, the vowel in the word-initial/pre-stress position ([тɐ́кápa] is longer than the vowel in the word-medial/post-stress position ([ákəta]). For random effect, we included random intercepts for participants and clusters. In our model selection process, we used backwards elimination of non-significant effects. The full model included NATIVENESS (CC vs. CVC) and WORD POSITION (word-initial vs. word-medial) as fixed factors, as well as the interaction between them. The second model excluded the interaction NATIVENESS × POSITION from the full model. Both Bayesian Information Criterion (BIC) values (Schwarz et al., 1978) and Akaike Information Criterion (AIC) values (Akaike, 1987) were used to assess model fit. *P*-values were generated using likelihood ratio tests.

For F1 and F2, multivariate regression models were performed using the MCMCglmm package in R. The independent variable in this model was NATIVENESS (CC vs. CVC). Moreover, the interaction between C1 TYPE (t, p, k) and NATIVENESS was included as an independent variable, because we observed in the earlier transcription task that the quality of vocoids between two consonants was affected by C1 type. GENDER (Male vs. Female) was also included as an independent variable, since gender differences in the formant frequency values of adults are well established (Peterson and Barney, 1952; Fant, 1966, 1975). For contrast coding of categorical variables, the binary variables, NATIVENESS and GENDER, were coded using sum-to-zero contrast coding such that each factor had a mean of zero. The 3-level factor, C1 TYPE, was coded so that each level

of the factor (C1-/t/, C1-/k/) was compared to the reference level (C1-/p/). We included random intercepts for participants and items. The critical independent variable in question was NATIVENESS, and for this variable, we included correlated random slopes for participants and clusters (this quantifies by-participant and by-cluster variability in the effect of NATIVENESS). We reported 95% highest posterior density (HPD) interval for each coefficient and related p-values.

4.3.2.1 Duration of inserted vocoids

Figure 4.5 shows the distribution of the vocalic duration of each inserted vocoid in CCs (inserted vocoids being characterized by voice bar, clear formants and higher intensity than flanking consonants) and each vowel in CVCs in word-initial position (top) and word-medial position (bottom). As shown in Table 4.4, the minimal duration of the inserted vocoid was 11.09ms in the production of /knapa/ and the maximal duration was 122.78ms in the production of /klapa/. The average in word-initial position was 37 ms (SD = 11.7) and in word-medial position was 38ms (SD = 12.9) . For the vowels in CVC, the minimal value was 12.13ms and the maximal value was 187.6ms. The average in word-initial position was 76ms (SD = 24.8) and in word-medial position was 44ms (SD = 15.1).

Table 4.4: Descriptive statistics on the duration (ms) of the vocoids in CCs and vowels in CVCs

Nativeness	Position	Mean	Std. Dev.	Range
CC	Word-initial (pre-stress)	37	11.7	11.09-122.78
	Word-medial (post-stress)	38	12.9	
CVC	Word-initial (pre-stress)	76	24.8	12.13-187.6
	Word-medial (post-stress)	44	15.1	

To understand the effect of NATIVENESS on the duration of vocoids, we conducted a series of two nested linear mixed-effects models for the vocoid duration data, as shown in Table 4.5.

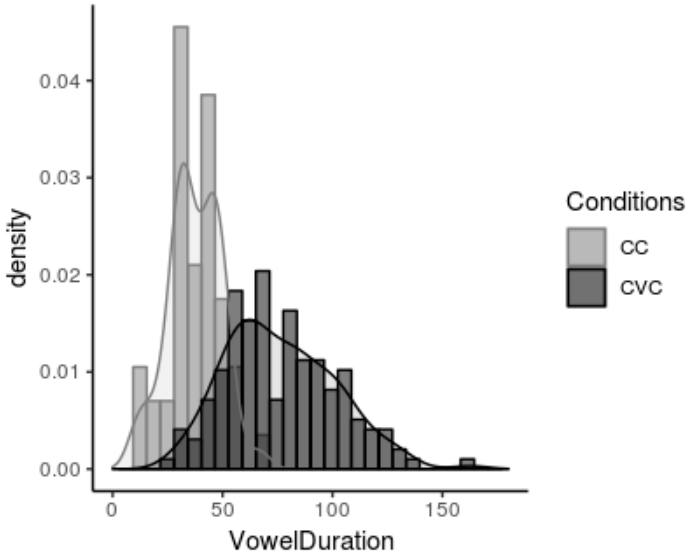
Table 4.5: Model comparison showing effect of nativeness and word position on vocoid duration. Models are given in parentheses. (N=Nativeness, P=Position, p=participant, c=cluster)

Model comparison	Df	AIC	BIC	logLik	Chisq	Pr(>Chisq)
$N+P+N\times P+(1\|p)+(1\|c))$	7	3423.8	3447.6	-1705.9		
$N\times P(N+P+(1\|p)+(1\|c))$	6	3423.8	3447.6	-1705.9	68.699	$<$ 2.2e-16

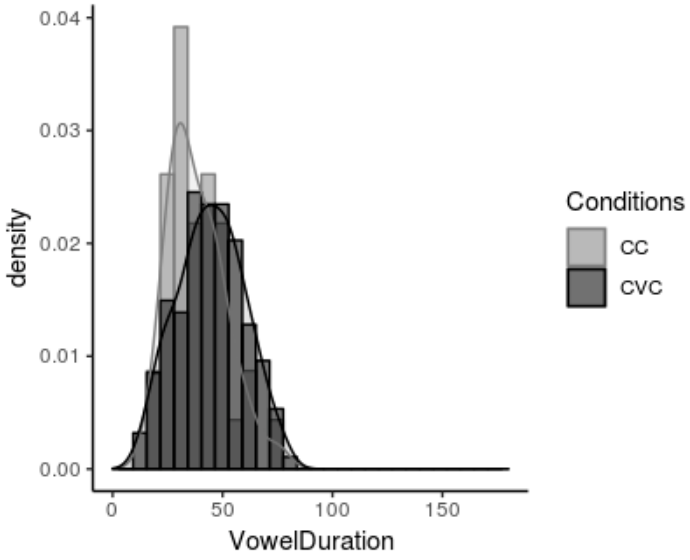
To summarize the model comparison, dropping the interaction between NATIVENESS and WORD POSITION and keeping only the independent variable WORD POSITION changed the model. Both BIC and AIC values for this model were the lowest. Significant interaction between NATIVENESS and WORD POSITION indicates that there was an effect of NATIVENESS on the duration of vocoids, but not in both positions. As shown in Figure 4.5, in word-initial position, the vocoids in CCs are much shorter than the vowels in CVCs. However, in word-medial position, there is no such difference between them. The difference across word positions may be affected by the input productions. We saw that in the Russian stimuli, the vowel in CVCs in word-initial position (pre-stress) is much shorter than that in word-medial position (post-stress) (word-initial: mean = 76ms, SD = 11.6; word-medial: mean = 46ms, SD = 11.8). Mandarin speakers may have been sensitive to these duration differences, and may have accordingly produced similar vowel durations. As mentioned above, the average vowel duration in word-initial position was 76ms (SD = 24.8) and in word-medial position was 44ms (SD = 15.1).

Overall, the duration of the inserted vocoid in the production of non-native CCs was similar to the duration of a vowel for clusters in post-stress position, but 48% shorter than a vowel in pre-stress position.

Figure 4.5: Distribution of the vocalic duration of each inserted vocoid in CCs and vowels in CVCs across word positions (word-initial (top); word-medial (bottom)). The grey line shows the duration of V in the production of the non-native target CCs; the black line shows the duration of V in the production of the CVC controls.



(a) Word-initial position



(b) Word-medial position

4.3.2.2 First and second formants (Vocoid quality)

Table 4.6: Descriptive statistics on the F1 and F2 values (Hz) of the vocoids

Formants	Mean	Std. Dev.	Range
F1	476	132	246–797
F2	1628	199	1092–2067

Table 4.6 shows descriptive statistics on the F1 and F2 values (Hz) of the vocoids between two consonants in the production of non-native CCs. The formant values indicate a schwa-like quality.

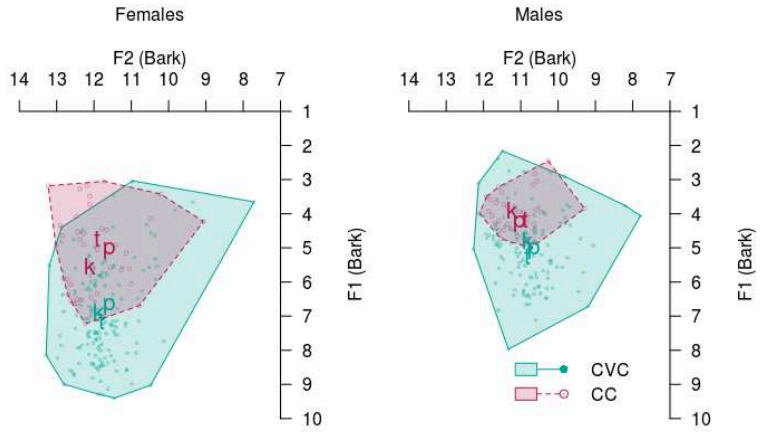


Figure 4.6: F1 and F2 of the inserted vocoids (CC) and vowels (CVC), with formant frequencies converted to Bark scale. Plotted F1 and F2 mean values for each C1 type (*p, t, k*). The polygonal boundary line is a property of the entire vocoid space (in pink) and vowel space (in green).

Figure 4.6 shows the F1 and F2 of the inserted vocoids and lexical vowels for each gender, with formant frequencies converted to Bark scale. The multivariate regression model shows that the overall F1 and F2 values of vocoids in non-native CCs did not differ from those of vowels in control CVC sequences ($CVC = -0.02$, *95% HPD* $[-3.46, 3.68]$, $p = 0.98$). That is, as shown in Figure 4.6, the vocalic space of the vowels 'covers' that of the vocoids. The F2 of the inserted vocoids in CCs did not differ from the F2 of the vowels in CVCs ($F2 \times CVC = -0.16$,

95% HPD $[-1.1, 0.73]$, $p = 0.72$). Only the F1 of the inserted vocoids in CCs was lower than that of the vowels in CVCs (*F1* $\times$ *CVC* $= 1.5$, *95% HPD* [0.47, 2.39], $p = 0.007$). Lower F1 values indicate a higher vowel, meaning that participants produced the vocoids with the mouth minimally open and the jaw or tongue body raised. In addition, recall that vocoids in word-initial clusters are in pre-stress position, and their quality is approximately [ɐ]. Mandarin speakers produced this vowel as an open vowel in CVCs sequences 47% of the time. The F1 of the vowel in CVC is significantly higher than the F1 of the vocoids in CCs.

An expected significant effect was the effect of GENDER, as both F1 and F2 of male speakers were lower than for female speakers (*F1* $\times$ *male* $= -1.8$, *95% HPD* $[-2.28, -1.31]$, $p < 6e-04$; *F2* $\times$ *male* $= -0.97$, *95% HPD* $[-1.35, -0.59]$, $p < 6e-04$, respectively).

C1 TYPE did not significantly affect the formant frequencies of the inserted vocoid in production. Instead, the vowel space was usually centralized to a schwa. This result is not consistent with our findings in the transcription task.

4.3.3 Duration of non-native CCs

In trying to understand the nature of the vocoid, the question we ask here is whether the presence of an acoustic vocoid will change the duration of the cluster as a whole. We observed that inserted vocoids can be as long as 122.78ms. If the vocoid is a phonological segment, the presence of such a segment in a CC cluster should increase the duration of the cluster. If it is a 'transition', the presence or absence of the vocoid will not influence the duration of the cluster. This prediction is based on Ridouane and Fougeron (2011), who argued, based on the absence of a duration difference, that the vocoids found in Tashlhyit consonant sequences are transitions, and not inserted vowel segments.

To answer this question, we compared the duration of CCs and CVC controls

across different productions of the clusters (with vocoid, with voicing, without vocoid) for stop-stop and stop-nasal clusters. Stop-liquid and liquid-stop clusters were removed from the analysis, since the dark liquid was either deleted or changed into a back close vowel [u] or [o]. In addition, given that CC duration was measured from the offset of the vowel preceding the cluster to the onset of the vowel following the cluster, the comparison was restricted to word-medial (post-stress) clusters.

Figure 4.7 shows the duration of entire CC clusters with the following productions: *with vocoid, with voicing, without vocoid*, across different clusters, as well as the duration of CVC sequences *with vowel*. We observed that the duration of the entire CC interval was very similar across outputs, either with or without a vocoid. The duration of the CVC interval (*with vowel*) was slightly longer than the duration of the CC interval across all outputs (*with vocoid, with voicing, without vocoid*).

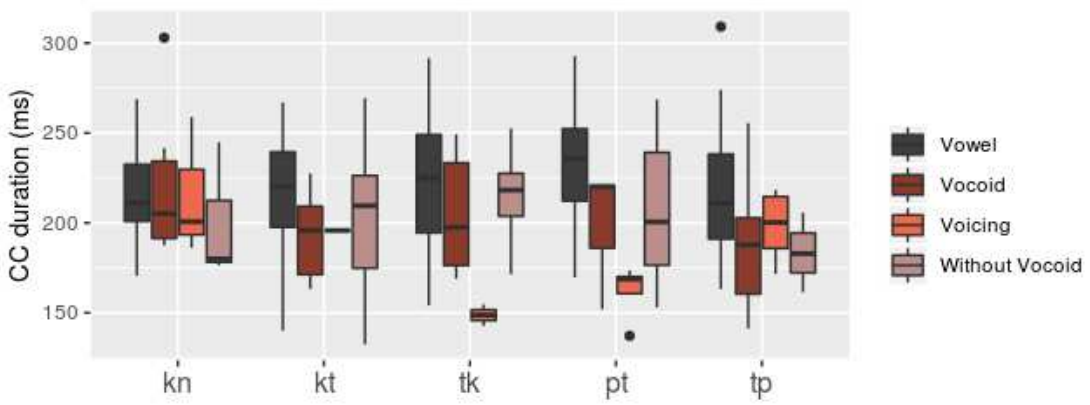


Figure 4.7: Duration of CVC with vowel and duration of CC with vocoid, with voicing, and without vocoid. Error bars indicate 95% bootstrap confidence intervals.

To assess the statistical significance of the observations in Figure 4.7 above, we fitted a series of four linear mixed-effect models to the CC duration data. The test fixed factor was OUTPUT (*with vocoid, with voicing, without vocoid, and with*

vowel). For this critical variable, we included random intercepts and slopes for participants as well as random intercepts for items.

In the model selection process, we conceptually separated the fixed effects into control variables and the test variable (OUTPUT). CLUSTER TYPE (*medial-*/kn/, *medial-*/kt/, *medial-*/tk/, *medial-*/tp/, *medial-*/pt/) was a control fixed factor. If it leads to a significant interaction with OUTPUT, this would indicate the effect of cluster types on acoustic timing lags across OUTPUT. The full model included OUTPUT (with vocoid, with voicing, without vocoid, and with vowel) and CLUSTER TYPE (*medial-*/kn/, *medial-*/kt/, *medial-*/tk/, *medial-*/tp/, *medial-*/pt/) as fixed factors, as well as the interaction between them. The second model excluded the interaction OUTPUT × CLUSTER from the full model. The third model excluded the fixed factor CLUSTER from the second model, to test the effect of cluster type. The fourth model excluded the fixed factor OUTPUT from the third model, to test the effect of output production type. Both Bayesian Information Criterion (BIC) values and Akaike Information Criterion (AIC) values were used to assess model fit. P-values were generated using likelihood ratio tests and Tukey's HSD post-hoc analyses using the emmeans package (Lenth, 2016).

As shown in Table 4.7, the model with OUTPUT as fixed factor was the best model to interpret the data. This is indicated both by a chi-squared test of the log-likelihood ratios, where $p = 0.00032$, and by the lowest AIC value and the lowest BIC value. There was a main effect of OUTPUT on the duration of the CC interval.

Tukey's HSD post-hoc tests further showed that there was no significant difference between the non-native CC with vocoid and CVC with vowel conditions ($p's > 0.05$), except for the cluster /tp/ ($\beta = 33, p = 0.0057$). Participants may have used the same timing patterns in producing the native CVC sequences and the non-native CC clusters with vocoids. The similar duration of CCs with vo-

Table 4.7: Model comparison showing effects of output and cluster type on the duration of CC interval. Models are given in parentheses. Tested fixed effects are given on the left-side of the model.

Model comparison	Df	AIC	BIC	logLik	Chisq	Pr(>Chisq)
(*Output + Cluster + Output × Cluster + (1 + Output /participant) + (1/item)*)	32	1867	1972	-901.46		
Output × Cluster (*Output + Cluster + (1 + Output/participant) + (1/item)*)	20	1864	1929	912	21	0.05
Cluster (***Output + (1 + Output/participant) + (1/item)***) (selected)	16	1859	1912	-914	3.5	0.48
Output (*1 + (1 + Output/participant) + (1/item)*)	13	1872	1914	-923	18.6	0.00032

coids and CVCs with vowels suggests that, at least with respect to duration, the vocoid between two consonants may be a segment, the same as a vowel in the CVC controls.

Regarding the production of non-native CCs specifically, CC duration was longer with a vocoid than without one, as indicated by the positive coefficients for /kn, kt, tk, pt/ (*vocoid- without vocoid*: $\beta = 12.6; \beta = 1.4; \beta = 11.8; \beta = 15.5$, respectively), but this difference was not significant ($p's > 0.05$). Results indicate that across all output types (*with vocoid, with voicing, without vocoid*), there was no significant difference in CC duration.

Can these results tell us something about the nature of the vocoid? Our prediction was that clusters produced with a vocoid would be longer than clusters produced without a vocoid if the vocoid is an inserted vowel segment. The results indicate a trend in this direction, but without reaching significance. At the same time, we find that clusters produced with a vocoid are comparable in duration to CVC sequences with a full lexical vowel.

The two results taken together suggest to us that overall, Mandarin speakers used similar CC timing patterns (in terms of the duration of the CC interval) in the production of CCs with vocoids and their CVC controls. They may have used, in both conditions, the only consonant-to-consonant timing that they know, namely the one they use in the CVC sequences.

However, CCs without vocoids, which are the cluster-like production of CCs, are not significantly shorter than CCs produced with a vocoid. This suggests that Mandarin speakers try to adjust their CC timing pattern when prompted by non-native CC clusters. They may try to match the input production as best as they can, and possibly do not always perceive a vowel within the cluster. This is as far as our interpretation can go, but we cannot draw from these results a reliable conclusion about the nature of the vocoid in non-native cluster production.

4.3.4 Temporal coordination in the production of non-native CCs

To further understand the temporal coordination patterns in the production of non-native consonant clusters by Mandarin speakers, we analyzed the inter-plateau interval (IPI) and the inter-burst interval (IBI). We compared the acoustic timing lags of CC clusters produced with vocoids to those of the CVC controls. We predicted the following:

- If Mandarin speakers use the CVC timing patterns of their native language to produce non-native CCs, as we proposed based on the duration comparison discussed above, we should find a significant correlation between the acoustic timing lags in the production of non-native CCs and the CVC controls.

Since acoustic measurements of IBI can only be made with stop-stop clusters, we confine ourselves to the following clusters: /kt, tk, pt/ in word-initial position, and /kt, tk, pt, tp/ in word-medial position.

For statistical analysis, linear mixed-effect models were computed using the *lme4* package and model selections in R for all the analysis of acoustic timing lags (IPI and IBI). The dependent variable was the acoustic timing lag (either IPI or IBI duration). The test independent variable was OUTPUT, with the three

types of productions of non-native CCs (with vocoid, with voicing, without vocoid) and CVC controls (with vowel). We included by-participant and by-item random intercepts. By-participant random slopes for OUTPUT were also included.

In the model selection process, we conceptually separated the fixed effects into control variables and the test variable (OUTPUT). CLUSTER TYPE (tk, kt, tp, pt) and WORD POSITION (initial vs. medial) were two control fixed factors.

If they lead to a significant interaction with OUTPUT, this would indicate that temporal coordination patterns in the production of non-native CCs could vary across cluster types and/or word positions. Acoustic timing lags in the INPUT were also included. A main effect of INPUT would indicate that the acoustic timing lags in the Russian stimuli affect the production of non-native CCs by Mandarin speakers. The full model included OUTPUT (with vocoid, with voicing, without vocoid, and with vowel), CLUSTER TYPE, WORD POSITION and acoustic timing lags in the INPUT as fixed factors, as well as the interactions among OUTPUT, CLUSTER TYPE and WORD POSITION. The second model removed the three-way interaction from the full model. The third model removed from the second model one of the two-way interactions, WORD POSITION × CLUSTER TYPE, which may have had less effect on the model. The fourth model removed the two-way interaction, OUTPUT × CLUSTER TYPE, from the third model. The fifth model removed the two-way interaction, OUTPUT × WORD POSITION, from the fourth model. The sixth model removed the fixed factor INPUT from the previous model. The seventh model removed the fixed factor CLUSTER TYPE from the previous model. The eighth model removed the fixed factor WORD POSITION from the previous model. The eighth model removed the fixed factor OUTPUT from the previous model. All models included random intercepts for participants and items. Both Bayesian Information Criterion (BIC) values and Akaike Information Criterion (AIC) values were used to assess model

fit. P-values were generated using likelihood ratio tests and Tukey's HSD post-hoc analyses using the *emmeans* package.

In addition, to determine whether Mandarin speakers used the timing patterns of their native language to produce non-native CCs, we tested the correlation between the timing patterns in the production of non-native CCs and the CVC controls, using Pearson's Correlation Coefficients.

4.3.4.1 Acoustic timing lag: inter-plateau interval (IPI)

Figure 4.8 shows the duration of the inter-plateau interval (IPI) in the production of non-native CCs with *vocoid*, with *voicing*, and *without vocoid* as well as in the production of the CVC control with a lexical *vowel* in the same consonantal contexts.

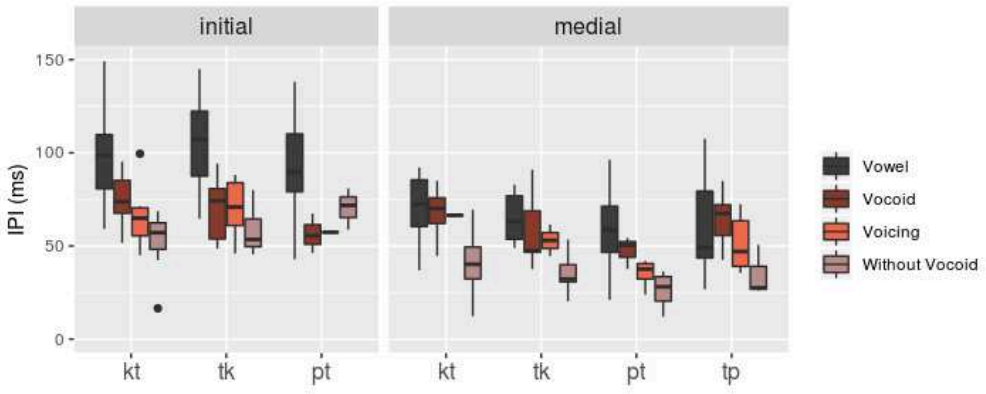


Figure 4.8: Duration of inter-plateau interval (IPI). Error bars indicate 95% bootstrap confidence intervals.

To test different IPI durations across different OUTPUT in both positions, linear mixed-effect models were computed. Table 4.8 shows the models listed from the most complex model to the most simple one, with the four fixed factors: OUTPUT, WORD POSITION, CLUSTER TYPE, and INPUT, and the interactions among OUTPUT, WORD POSITION, and CLUSTER TYPE.

The results indicate an interaction between OUTPUT and WORD POSITION, as removing this predictor significantly affected the model fit. The best model supported by the data was the one with four fixed factors and the interaction between Output and Cluster Type. This decision was made based on both the significance of the difference in log-likelihood between the third and the fourth model through a chi-square test, and the lowest AIC and BIC values. This effect is illustrated in Figure 4.8. OUTPUT had a significant effect on IPI that was different across word positions ($\chi^2(21) = 16.8, p = 0.00078$). Recall that vowels in CVCs in word-initial/pre-stress position were longer than in word-medial/post-stress position. Since the IPI included C1 release and vocoid (if present) or vowel, the IPI of CVCs *with vowel* in word-initial position was longer than that in word-medial position. In addition, Tukey's HSD post-hoc analyses indicate that the IPI *with vocoid and with voicing* in word-medial position was longer than the IPI *without vocoid* ($p < 0.01$). Interestingly, in word-initial position, there was no significant difference. This means that participants used similar IPI patterns to produce CCs in word-initial position, with or without a vocoid. The average IPI duration *without vocoid* was 57.66ms (SD = 14.21) in word-initial position, and the average duration *with vocoid* was 70.77ms (SD = 15.43).

There was also a significant effect of CLUSTER TYPE ($\chi^2(17) = 8.05$, $p = 0.045$). Tukey's HSD post-hoc analyses indicate that IPI was shorter for /pt/ than for /kt/ and /tk/ across outputs (*with vowel, with vocoid, with voicing, without vocoid*) in both positions (p's < 0.05). For /tp/, there was no significant difference from /pt/. Acoustic timing lags in the INPUT Russian stimuli may have influenced the IPI duration, as indicated by the chi-square test of goodness of fit ($\chi^2(20) = 15.8$, $p < 0.0001$). Participants produced similar IPI durations to what they had heard. As shown in Figure 4.3, in the Russian stimuli, the IPI for /pt/ was shorter than for /kt/ and /tk/. Another explanation for the timing

difference is that it may be related to articulator-specific patterns. The cluster /pt/ is produced with different articulators, namely the upper lip, lower lip, and jaw for the labial /p/, and the tongue tip for the alveolar /t/. The movement from one gestural target to another should thus be faster than when only one articulator is used to reach two different targets, as is the case in the cluster /kt/, where the same articulator, the tongue, is used for the velar and for the alveolar target constrictions. If this is true, then the timing lag for /pt/ should be shorter than it is for /kt/, and this articulator pattern may be copied in the production of non-native speech.

4.3.4.2 Acoustic timing lag: inter-burst interval (IBI)

Figure 4.9 shows the duration of the inter-burst interval (IBI) in the production of non-native CCs with *vocoid*, with *voicing*, and *without vocoid* as well as in the production of the CVC controls with a lexical *vowel* in the same consonantal contexts.

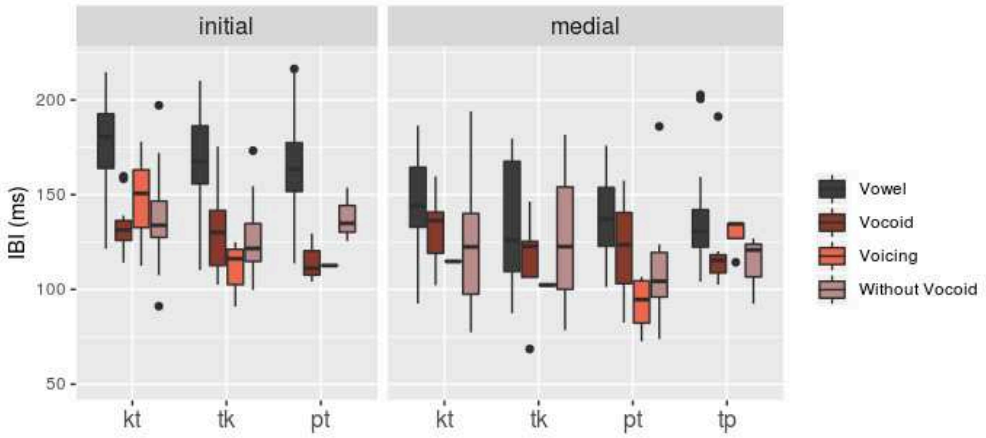


Figure 4.9: Duration of the inter-burst interval (IBI). Error bars indicate 95% bootstrap confidence intervals.

Table 4.9 shows statistical analyses of the trends displayed in Figure 4.9. The

Table 4.8: Model comparison showing effect of output, word position, and cluster type on IPI duration. Models are given in parentheses.

Model comparison	Df	AIC	BIC	logLik	Chisq	Pr(>Chisq)
(*Output + Cluster Type + Position + Cluster Type × Position + Output × Cluster Type + Output × Position + Output × Cluster Type × Position + Input+ (1+Output|participant) + (1|item)*)	41	3007	3165	-1463		
Output × Cluster Type × Position (*Output + Cluster Type + Position + Cluster Type × Position + Output × Cluster Type + Output × Position + Input+ (1+Output|participant) + (1|item)*)	35	2999	3133	-1465.5	3.2	0.7826
Cluster Type × Position (*Output + Cluster Type + Position + Output × Cluster Type + Output × Position + Input+ (1+Output|participant) + (1|item)*)	33	3001	3128	-1467	5.8	0.0544
Output × Cluster Type (***Output + Cluster Type + Position + Output × Position + Input + (1+Output|participant) + (1|item)***) (selected)	24	2996	3088	-1473.8	12.7	0.17
Output × Position (*Output + Cluster Type + Position + Input +(1+Output|participant) + (1|item)*)	21	3006	3087	-1482	16.8	0.00078
Input (*Output + Cluster Type + Position + (1+Output|participant) + (1|item)*)	20	3020	3097	-1490	15.8	6.898e-05
Cluster Type (*Output + Position+ (1+Output|participant) + (1|item)*)	17	3022	3087	-1494	8.05	0.045
Position (*Output + (1+Output|participant) + (1|item)*)	16	3053	3115	-1511	33	1.391e-06
Output (*1 + (1+Output|participant) + (1|item)*)	13	3077	3127	-1525	30	1.391e-06

same models were computed for the IBI and IPI. Results indicate a significant interaction between output and word position for IBI duration, as removing this predictor significantly affected the model fit. The best model supported by the IBI data was the one with four fixed factors and this interaction. This decision was made based on both the significance of the difference in log-likelihood between the third and the fourth model through a chi-squared test, and the lowest AIC and BIC values. This effect is illustrated in Figure 4.9, which shows that OUTPUT had an effect on IBI duration that was somewhat different across positions ($\chi^2(21) = 13.07, p = 0.0045$).

However, IBI results were not consistent with the IPI results, as there was less difference among the outputs. Tukey's HSD post-hoc analyses indicate that in word-initial position, the IBI *with vowel* (i.e. in control CVC) was longer than the IBI of clusters produced *with voicing* ($\beta = 41, p < 0.032$), but was not different from the other outputs. However, in word-medial position, there was no difference among the outputs. Participants used similar IBI patterns to produce non-native CCs (with or without a vocoid) and CVC sequences in both word-initial and word-medial position.

There was, again, a significant effect of CLUSTER TYPE ($\chi^2(17) = 8.4$, $p = 0.038$). As with IPI, Tukey's HSD post-hoc analyses indicate that IBI was shorter for *pt* than for *kt* across outputs (*with vowel, with vocoid, with voicing, without vocoid*) in both positions ($p's < 0.05$). For *tp* and *tk*, there was no significant difference from *pt*. This finding is consistent with the IPI results, which showed that the IPI for /pt/ was shorter than for /kt/ and /tk/. One explanation for this difference is that participants may seek to reproduce what they perceive, including such fine-grained acoustic details as those just discussed. Acoustic timing lags in the INPUT Russian stimuli also influenced IBI duration, as indicated by the chi-square test of goodness of fit ($\chi^2(20) = 10.11$, $p = 0.0015$). As shown in Figure

4.3, in the Russian stimuli, the IBI for /pt/ was shorter than for /kt/ in both positions. As discussed above for the IPI, another explanation for the timing difference may be that it is due to articulator-specific patterns. The cluster /pt/ is produced with different articulators, whereas /kt/ involves the same articulator for both target constrictions. Overall, the timing lag for /pt/ should be shorter than it is for /kt/. If this is true, we suggest that even though both /pt/ and /kt/ are illicit for Mandarin speakers, the articulator-specific patterns may influence their non-native production.

4.3.4.3 Correlation between IPI and IBI across productions

We have seen so far that the IPI and IBI of the acoustic timing lags varied in the production of CCs and CVCs. Further, the two acoustic timing lags were not consistent with each other across Outputs (*with vowel, with vocoid, with voicing, without vocoid*).

We want to know whether Mandarin speakers use the same articulatory timing pattern in the production of non-native CC clusters as in CVC controls. We are interested in verifying a correlation between IBI duration and vocoid/vowel duration on one hand, and IPI duration and vocoid/vowel duration on the other hand. We compare two conditions: in CC clusters and CVC controls. We also compare the correlation between IBI and IPI across the possible production outputs (*with vowel, with vocoid, with voicing, without vocoid*). We predict the following:

- if speakers are using the same articulatory timing patterns in both non-native CCs and CVC controls, the timing patterns should be positively correlated in both conditions. Contrary findings would suggest that speakers produce the non-native clusters using a different articulatory timing pattern from the native one.

Table 4.9: Model comparison showing effect of output, word position, and cluster type on IBI duration. Models are given in parentheses.

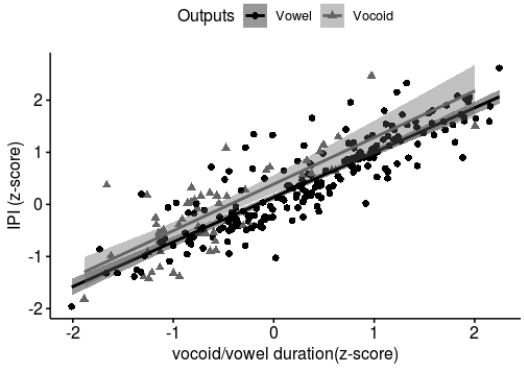
Model comparison	Df	AIC	BIC	logLik	Chisq	Pr(>Chisq)
(*Output + Cluster Type + Position + Cluster Type × Position + Output × Cluster Type + Output × Position + Output × Cluster Type × Position + Input+ (1+Output/participant) + (1/item)*)	41	3189	3347	-1554		
Output × Cluster Type × Position (*Output + Cluster Type + Position + Cluster Type × Position + Output × Cluster Type + Output × Position + Input+ (1+Output/participant) + (1/item)*)	35	3182.3	3316.7	-1556.2	4.6526	0.59
Cluster Type × Position (*Output + Cluster Type + Position + Output × Cluster Type + Output × Position + Input+ (1+Output/participant) + (1/item)*)	33	3178.5	3305.1	-1556.2	0.1256	0.9391
Output × Cluster Type (***Output + Cluster Type + Position + Output*** × ***Position + Input + (1+Output/participant) + (1/item)***) (selected)	24 3167.1	3259.2	-1559.6	6.6658	0.6719	
Output × Position (*Output + Cluster Type + Position + Input +(1+Output/participant) + (1/item)*)	21	3174.2	3254.8	-1566.1	13.07	0.0045
Input (*Output + Cluster Type + Position + (1+Output/participant) + (1/item)*)	20	3182.3	3259.1	-1571.2	10.11	0.0015
Cluster Type (*Output + Position+ (1+Output/participant) + (1/item)*)	17	3184.7	3250	-1575.4	8.4	0.038
Position (*Output + (1+Output/participant) + (1/item)*)	16	3204.8	3266.3	-1586.4	22.12	2.553e-06
Output (*1 + (1+Output/participant) + (1/item)*)	13	3224.3	3274.2	-1599.1	25.4	1.267e-05

4.3.4.4 Comparing CVC controls and CCs produced with vocoids

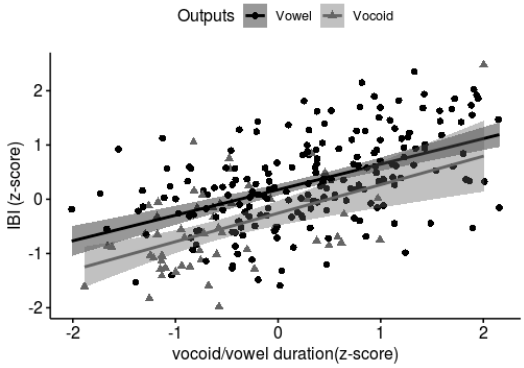
Mandarin speakers produced a vocoid after C1 30% of the time overall for stop-stop clusters. Figure 4.10 shows a scatter plot of IPI duration and vocoid/vowel duration (a) and a scatter plot of IBI duration and vocoid/vowel duration (b). Since we are comparing positions that have different average IPI and IBI durations, we z-scored IPI and IBI durations and vocoid/vowel durations within positions, to avoid obtaining spurious correlations (driven by differences across positions). The black dots represent tokens with vocoids inserted into non-native CCs, while the gray triangles are tokens produced with lexical vowels in CVC controls. The lines represent linear regression fits to the tokens with vocoids and lexical vowels in the production of CC and CVC, respectively.

To decide which correlation method can be used, we conducted Shapiro-Wilkes normality tests in R to verify data distribution. Results indicate that some data distributions were significant deviations from normality (where $p < 0.05$) while some data were normally distributed (where $p > 0.05$) (IPI_{vocoid}: $w = 0.97$, $p = 0.21$; IPI_{vowel}: $w = 0.99$, $p = 0.03$; IBI_{vocoid}: $w = 0.96$, $p = 0.09$; IBI_{vowel}: $w = 0.98$, $p = 0.015$; Vocoid: $w = 0.90$, $p = 0.0001$; Vowel: $w(227) = 0.96$, $p = 0.72$). Thus, we computed non-parametric Spearman correlation analyses to evaluate the nomotonic relationship between continuous variables. The Spearman correlation coefficient was interpreted as follows: 0.1–0.3: weak correlation; 0.4–0.6: moderate correlation; 0.7–0.9: strong correlation; 1: perfect correlation (Dancey and Reidy, 2008).

As shown in Figure 4.10, there was a significant positive correlation between IPI and vowel duration in the production of CVC controls ($\rho = 0.86$, $p < 2.2e-16$) as well as between IPI and vocoid duration in the production of non-native CC clusters ($\rho = 0.71$, $p <$ 2.2e-16). As vowel or vocoid duration increases, the duration of the IPI interval also increases. To test the difference between the two



(a) The correlation between IPI duration (y-axis) and vocoid/vowel duration (x-axis) across all stop-stop clusters in both positions.



(b) The correlation between IBI duration (y-axis) and vocoid/vowel duration (x-axis) across all stop-stop clusters in both positions.

Figure 4.10: Correlation between (a) IPI duration and Vocoid duration; (b) IBI duration and vocoid/vowel duration. Black dots represent the production of CCs with inserted vocoids. Gray triangles represent the production of CVCs with lexical vowels. The linear regression line fitted to the black dots shows a significant positive correlation; the gray triangles show a trend in the same direction. The gray bands around the line represent the 95 % confidence interval of the regression line.

correlation coefficients, Spearman correlations were treated as Pearson coefficients and Fisher Z-transformations were used (Myers and Sirois, 2006). Even though both coefficients displayed strong positive correlation, a Fisher Z-transformation analysis showed that there was a significant difference between them (*Fisher Z-transformed difference*: 2.6; $p = 0.0094$). For IBI, there were both moderate positive correlations between IBI and vowel duration in production of the CVC control and between IBI and the vocoid duration in non-native CC clusters ($\rho = 0.48$, $p < 2.2e{-}16$; $\rho = 0.46$, $p = 0.00039$, respectively). A Fisher Z-transformation analysis shows that there was no significant difference between the correlation coefficients (*Fisher Z-transformed difference*: 0.18; $p = 0.86$). Therefore, the relationship between the IBI duration and the vocalic elements in the production of CVC and of CC was uniform. However, the relationship between IPI duration and the vocalic elements was not consistent. These results indicate that the organization of the consonantal timing and vocalic elements in the production of non-native CCs was not always consistent with the native CVC patterns, and some interesting differences emerge. Even though in both CCs and CVCs, IPI duration was highly positively correlated with the duration of vocalic elements, vocoids were less strongly correlated with the IPI than vowels. The duration of vocoids was more constant than the duration of lexical vowels. Therefore, increased IPI duration can be attributed to increased duration of vowels in CVC, but not always to increased duration of vocoids in CC clusters.

Figure 4.11 shows a scatter plot of the IPI and IBI in the production of CCs with vocoids and CVCs with vowels. There was a moderate correlation between the IPI and IBI in the production of CVC controls ($\rho = 0.6, p < 2.2e{-}16$). There was a similarly moderate correlation between the two acoustic timing lags in the production of CCs with inserted vocoids ($\rho = 0.72$, $p < 2.2e\text{-}16$). Differences between the two coefficients were not significant (*Fisher Z-transformed difference*:

-1.42; $p = 0.15$). Thus, the acoustic timing lags of IPI and IBI in the production of CCs with inserted vocoids were similar to those in the production of the CVC control sequences.

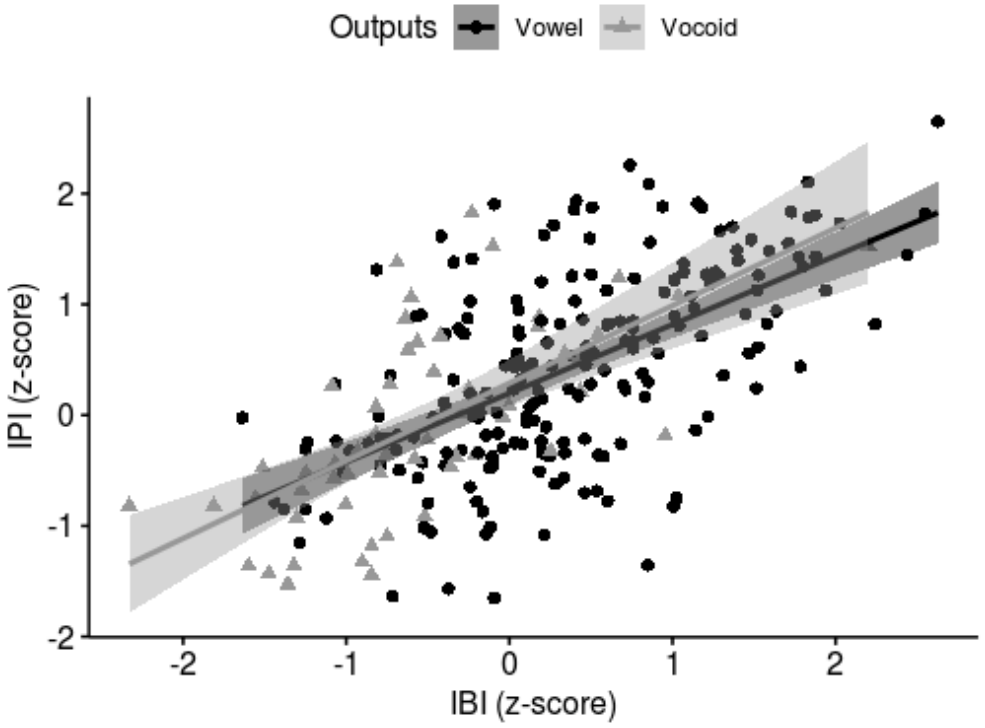


Figure 4.11: Correlation between IPI duration (y-axis) and IBI (x-axis) across all stop-stop clusters in both positions. Black dots represent the production of CC with vocoids. Gray triangles represent the production of CVCs with vowels. The linear regression line fit to the black dots shows a significant positive correlation; the gray triangles show a trend in the same direction. The gray bands around the line represent the 95 % confidence interval of the regression line.

Speakers used a similar articulatory timing pattern in the production of non-native CCs with vocoids and native CVCs with vowels. As regards the IPI, however, the duration of vocoids did not change the way it did for vowels. The duration of vowels was highly correlated with IPI duration. Increasing vowel/vocoid duration led to an increase in IPI duration, but this effect was less strong in the case

of vocoids. As mentioned before, IPI duration was measured from the onset of the release burst of C1 to the target achievement of C2, which included C1 release and the following vocalic elements (when present). If the vocoids in CC remain relatively constant in the IPI, the duration of the preceding C1 release should be positively correlated with the IPI.

Figure 4.12 shows the duration of C1 release in the following conditions: production of the non-native CC preceding inserted vocoids (black bars) and production of the CVC control preceding vowels (gray bars). We see that C1 release is indeed longer preceding inserted vocoids than preceding lexical vowels (Vocoid: mean = 29.37ms, SD = 13.69; Vowel: mean = 19.83ms, SD = 10.99).

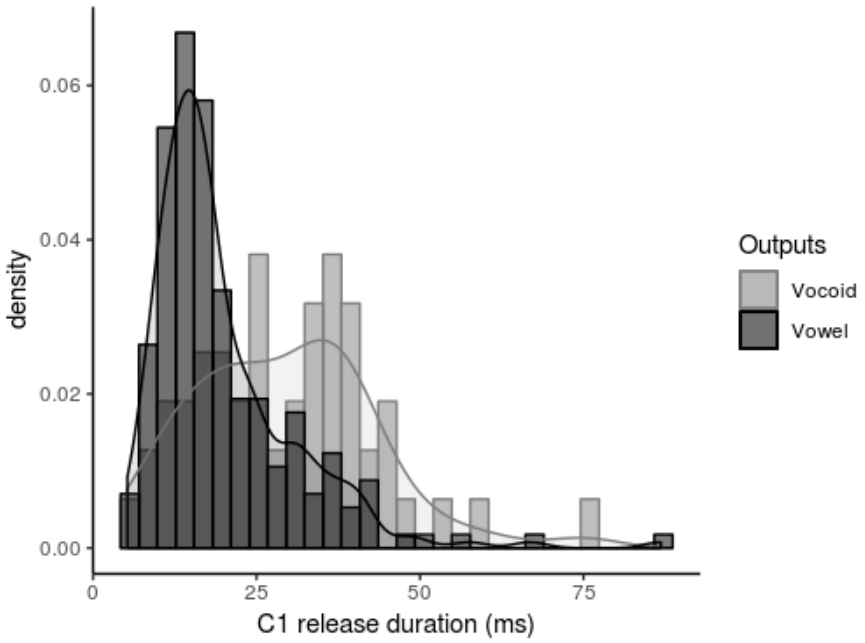


Figure 4.12: C1 release duration in the following conditions: production of the non-native CC with an inserted vocoid (black bars, mean = 29.37ms, SD = 13.69) and production of the CVC control with a vowel (gray bars, mean = 19.83ms, SD = 10.99)

Figure 4.13 shows the correlation between duration of C1 release and IPI. As in the previous scatterplots, we have z-scored C1 release duration within positions.

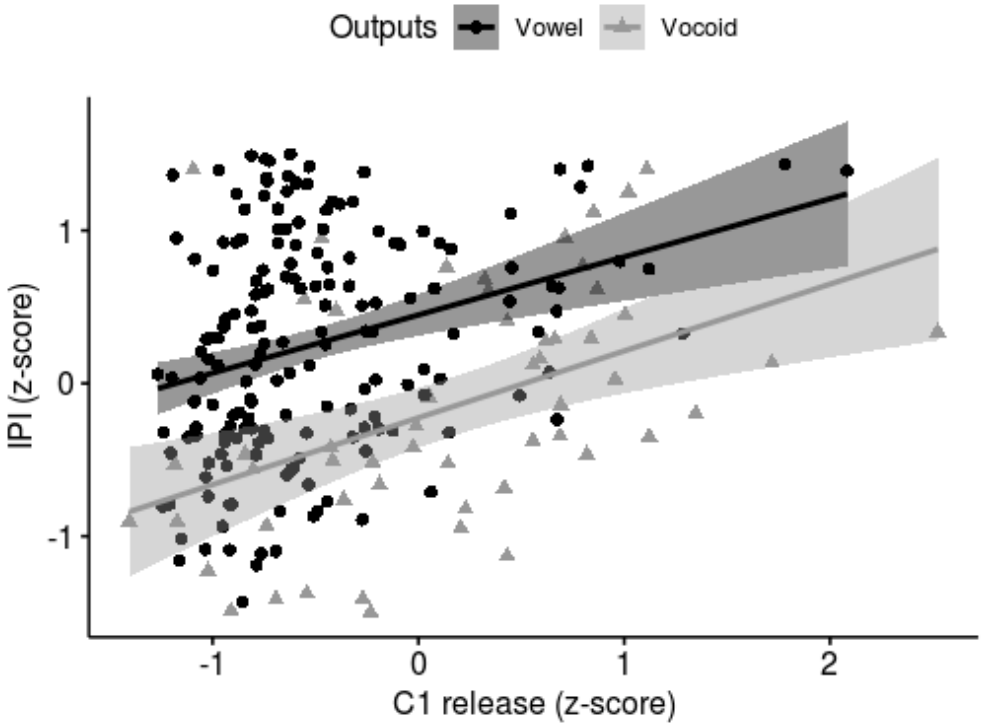


Figure 4.13: Correlation between IPI duration (y-axis) and C1 release duration (x-axis) across all stop-stop clusters in both positions. Black dots represent the two durations in the production of CC with inserted vocoids. Gray triangles represent the two durations for the production of CVC with vowels. The linear regression line fit to the black dots shows a significant positive correlation; the gray triangles show a trend in the same direction. The gray bands around the line represent the 95% confidence interval of the regression line.

The black dots represent tokens in the production of non-native CCs with the vocoids. The gray triangles are tokens in the production of CVC controls with the vowels. The lines are linear regression fits to the tokens in the production of CCs with vocoids and CVCs with vowels, respectively. C1 release preceding inserted vocoids was normally distributed, but not preceding vowels ($C1_{vocoid}$: $w = 0.97$, $p = 0.27$; $C1_{vowel}$: $w = 0.84$, $p =$ 1.979e-13). Non-parametric Spearman correlation analyses indicated there was a significant but weak positive correlation between IPI and C1 release duration in the production of the CVC controls ($\rho = 0.40$, $p =$

3.431e-09), but a stronger positive correlation between IPI and C1 release duration in the production of the non-native CCs ($\rho = 0.46$, $p = 0.00041$).

Overall, Mandarin speakers maintained the native CVC timing pattern in the production of non-native CCs with vocoids. Even though they were able to sometimes suppress the vocoid acoustically, they did not seem to adjust their articulatory timing. Instead, they retained the native timing pattern by increasing the release duration of the preceding consonant.

4.3.4.5 Comparing CVC controls and CCs produced with voicing

Mandarin speakers used a variety of modifications to adapt non-native consonant clusters. In addition to inserting an vocoid, they also produced a period of voicing after C1 (9% of the time overall for stop-stop clusters). Figure 4.14 compares the acoustic timing lag in CVC controls to that in CC clusters produced with voicing. The black dots represent tokens with vowel in the production of the CVC controls. The gray triangles are tokens with voicing in the production of non-native CC clusters. The lines are linear regression fits to the tokens with vowel and with voicing in the production of CVCs and CCs, respectively. Since some data distributions were significant deviations from normality (IPI: $w = 0.95$, $p = 0.24$; IBI: $w = 0.96$, $p = 0.41$), non-parametric Spearman correlation analyses were used. Recall that there was a moderate correlation between the IPI and IBI in the production of CVC controls ($\rho = 0.6$, $p <$ 2.2e-16). However, there was no correlation between the IPI and IBI in the production of CCs with voicing ($\rho = 0.23$, $p = 0.27$). In the production of voicing, increasing IBI duration did not increase IPI duration. It should be noted that the sample size for voicing was very limited, so that more data will need to be collected in the future to verify the findings of the present study. The main conclusion to be drawn here is that unlike in the production of CCs with vocoids, speakers did not use native CVC timing

patterns in the production of CCs with voicing.

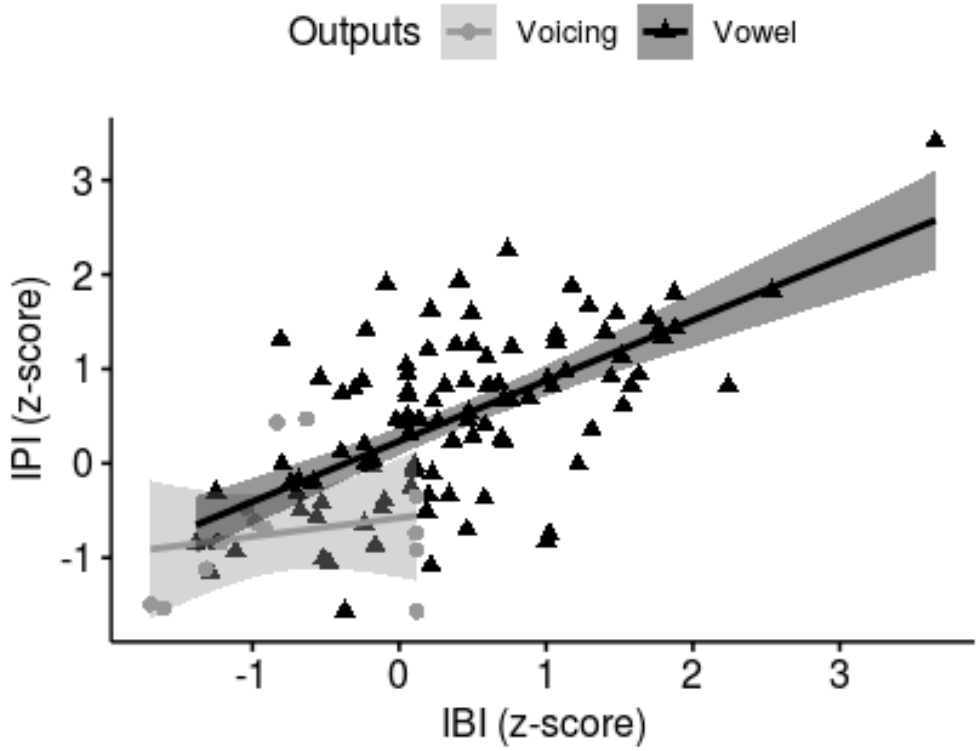


Figure 4.14: Correlation between IPI duration (y-axis) and IBI (x-axis) across all stop-stop clusters in both positions. Black dots represent the production of CC with voicing. Gray triangles represent the production of CVCs with vowels. The linear regression line fit to the black dots shows a significant positive correlation; the gray triangles show a trend in the same direction. The gray bands around the line represent the 95 % confidence interval of the regression line.

4.3.4.6 Comparing CVC controls and CCs produced without vocoids

Mandarin speakers produced non-native CCs without a vocoid between two consonants 20% of the time. Figure 4.15 compares the acoustic timing lag in CVC controls to that in CC clusters produced without a vocoid. The black dots represent tokens with vowels in the production of the CVC controls. The gray triangles are tokens without vocoid in the production of non-native CC clusters. The lines

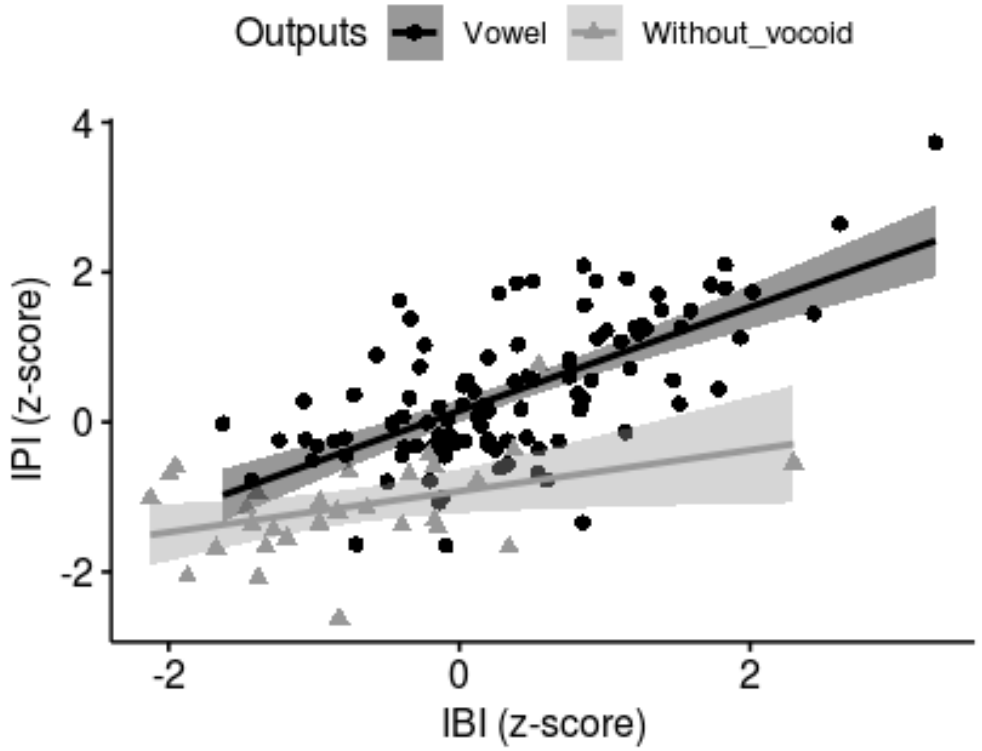


Figure 4.15: Correlation between IPI duration (y-axis) and IBI (x-axis) across all stop-stop clusters in both positions. Black dots represent the production of CCs without vocoids. Gray triangles represent the production of CVCs with vowels. The linear regression line fit to the black dots shows a significant positive correlation; the gray triangles show a trend in the same direction. The gray bands around the line represent the 95 % confidence interval of the regression line.

are linear regression fits to the tokens with voicing and without vocoid in the production of CCs, respectively. Since some data distributions were significant deviations from normality (IPI: $w = 0.98$, $p = 0.35$; IBI: $w = 0.94$, $p = 0.01$), non-parametric Spearman correlation analyses were used. The results indicated that there was significant but weak correlation between IPI and IBI in the production of CCs without a vocoid ($\rho = 0.31$, $p = 0.02$). In the production of without vocoid, increasing IBI duration did increase IPI duration. However, this correlation is weaker than the correlation between the IPI and IBI in the production of CVC

controls. Recall that there was a moderate correlation between the IPI and IBI in the production of CVC controls ($\rho = 0.6$, $p < 2.2\text{e-}16$). A Fisher Z-transformed difference analysis indicated that there was no difference between the two correlation coefficients (*Fisher Z-transformed difference*: 1.73; $p = 0.08$). Unlike in the production of CCs with vocoids, timing patterns in the production of CCs without vocoids used a different strategy from the native CVCs timing pattern.

4.4 Interim summary

To summarize, Mandarin speakers produced a vocoid between the two consonants of the cluster 26% of the time, all types of clusters taken together. Such vocoids are identified by the presence of a voice bar, clear formants, and higher intensity than the flanking consonants. Their duration is similar to the duration of unstressed vowels in CVCs sequences in word-medial position, but shorter than that of vowels in word-initial position. The F1 and F2 values of the vocoids are centralized to a schwa-like quality. In addition, the duration of the CC cluster with a vocoid is similar to the duration of a CVC sequence with a full lexical vowel. All these acoustic characteristics are similar to those of a reduced vowel in Mandarin, as reported by (Lin and Yan, 1980).

In addition to vocoid insertion, Mandarin speakers also produced the target clusters with consonant deletion and feature change. We found that C1 change was triggered by low amplitude of C1, a result which is consistent with the findings of our perception experiment. Interestingly, Mandarin speakers produced CC clusters without any modifications 13% of the time, and with only a voice bar between the two consonants 7% of the time. Correct production without a vocoid indicates that Mandarin speakers were sensitive to the phonetic details in the inputs and may even have 'correctly' perceived non-native consonant clusters. Monolingual

speakers, who had never heard these consonant clusters before, faithfully decoded the phonetic details in the Russian stimuli by relying on the auditory inputs.

As for the gestural coordination patterns of the possible productions—with vocoid, without vocoid and with voicing—the pattern for production with a vocoid was consistent with that for production with a vowel in CVC sequences. However, the patterns for production without a vocoid and with voicing were different from the articulatory patterns in native CVC sequences. These findings suggest a considerably high degree of variability in gestural coordination, showing that Mandarin speakers use various strategies to produce non-native consonant clusters.

Chapter 5

Discussion and conclusion

The current study investigated the perception and production of non-native consonant clusters by monolingual speakers of Mandarin Chinese. Mandarin exhibits moderately complex syllable structure, as discussed in Section 1.4, the most complex syllable type being CGVX (where C = consonant, G = glide, V = vowel, and X = nasal or off-glide). The permitted consonant sequences in Mandarin are #CG in word-initial position, and N#C in word-medial position. We wanted to know to what extent the production and perception of non-native CC sequences are influenced by phonological knowledge of the native language (phonotactic knowledge and gestural coordination patterns specific of the native language), and to what extent it is influenced by sensitivity to phonetic details (such as duration and intensity of the burst release). To answer these questions, we carried out several experiments, which included perceptual discrimination and production prompted by an audio model. We begin with a short summary of the experiments.

In an ABX discrimination test, Mandarin speakers were asked to discriminate between non-native (Russian) CC clusters and control CVC sequences (e.g., /ptaka/ vs. /pɐtaka/). We hypothesized that if native phonotactic knowledge primarily affects perception, Mandarin speakers would always perceive an epenthetic

vowel between the two consonants of the cluster, and would not be able to correctly discriminate between CC and CVC. If phonetic details in the non-native stimuli primarily contribute to non-native perception, then Mandarin speakers would not always perceive a vowel in the cluster, and would be at least partly able to discriminate between CC and CVC, specifically when no phonetic details are present that can be interpreted as a vowel.

A separate perception experiment was conducted, consisting of a transcription task, in order to verify what additional perceptual modifications may arise (e.g., C deletion, C feature changes), when hearing CC clusters.

Finally, a prompted production task, in which Mandarin speakers heard and repeated Russian stimuli, was used to investigate the production of non-native consonant clusters. In analyzing speakers' production, we focused on identifying the effect of native gestural coordination patterns on the production of non-native CCs through the analysis of acoustic properties of the vocoids [1] and measures of consonant-to-consonant acoustic timing lag. Coordination patterns are considered here to be language-specific, and thus part of phonological knowledge, consistently with the view taken by Articulatory Phonology (Browman and Goldstein, 1986, 1992a)

In the following sections, we discuss our results in relation to the questions and predictions presented in Section 1.5. We then return to the central question of this dissertation: how do native speakers of languages with simple or moderately complex syllable structure perceive and produce non-native clusters (CCs), a more complex type of syllable structure?

[1] Recall that in production, these acoustic vowels have received two different treatments. Under one view they are considered epenthetic vowels; under an other view, they are considered acoustic transitions between consonants. The two types of vowels may even be mixed in the production of non-native CCs (Shaw and Davidson, 2011). Thus, we use a phonetic terminology 'vocoid' to refer to such vowels in the case of production of non-native consonant clusters.

5.1 The role of phonotactic knowledge

We wanted to know to what extent the production and perception of non-native CC sequences can be influenced by phonotactic knowledge of the native language. It was predicted that, if native phonotactic knowledge drives perception, native speakers of Mandarin will systematically perceive vowels in the non-native clusters, and will also systematically produce them. The results are not consistent with this prediction, and show a mismatch between perception and production.

In the ABX discrimination task, results show that the sensitivity of Mandarin speakers to the phonotactic contrast was significantly above chance. Mandarin speakers misperceived the difference between non-native stop-stop (SS) clusters (e.g., ptaka) and the stop-vowel-stop (SVS) controls (e.g., pɐtaka) only 35% of the time. Based on this first result we can already conclude that native phonotactics alone does not drive non-native perception. In the follow-up transcription experiment, we tested more clusters including stop-stop (SS), stop-liquid (SL), stop-nasal (SN), and liquid-stop (LS). Mandarin speakers transcribed a vowel between two consonants in 62% of the total responses. They also sometimes deleted one of the consonants, or changed a feature of the consonant. The percentage of vowel transcriptions is unexpectedly high compared to the discrimination results. We believe that the increase of vowels in the transcription task may be due, as mentioned earlier, to the bias of Pinyin orthography, where a vowel between two consonants is obligatory. Therefore, participants may have transcribed a vowel even when they did not perceive one. To compensate for the effect of orthography and focus on the decoding of fine-grained phonetic properties in the inputs, we carried out a prompted production experiment. We found that, when hearing and repeating a Russian stimulus, Mandarin speakers produced an vocoid between two consonants of the cluster 26% of the time, all types of clusters taken together. Such

vocoids are identified by the presence of a voice bar, clear formants, and higher intensity relative to the flanking consonants. This is a relatively low percentage, which supports the initial conclusion that non-native perception is not entirely driven by native language phonotactics.

So far we have relied on data that informs us of the presence or absence of an epenthetic vowel (or a vocoid) in the cluster. But Durvasula et al. (2018) claimed that, in addition to the frequency of vowels, the quality of the vowels may be influenced by the phonotactics of the native language. In the current study, such an effect was only confirmed in the transcription experiment, but not in the prompted production experiment. In the transcription experiment, we found that the epenthetic vowel followed by a velar or an alveolar stop was transcribed with <e> in Pinyin, e.g., /ktapa/ → <ketapa>. In Pinyin, the symbol <e> stands for a mid back rounded vowel [ɤ]. The choice of [ɤ] may be due to the fact that it is an allophone of the mid vowel /ə/, which is the neutral vowel in Mandarin Chinese, and can be deleted in fast conversational speech, preceding a nasal consonant (Weinberger, 1997). This vowel could best approximate a transition between two consonants in a cluster. However, [ɤ] was not transcribed after a labial consonant. When preceded by a labial stop, the vowel was transcribed with <u> in Pinyin (e.g., /ptaka/ → <putaka>). This transcription choice is consistent with phonotactic restrictions in Mandarin (Lin, 2007), since in this case, the vowel agrees with the preceding labial consonant. This result is also consistent with the findings of the loanword study by Miao (2005), where the labial vowel /u/ is always inserted after the labial consonant /p/. Nevertheless, in spite of the consistency in transcription, the vocoid produced in prompted production after different C1—/p, t, k/—were similar. The quality of these vocoids was a centralized schwa. Based on these results, we claim that phonotactic knowledge is not the only factor affecting the perception and production of non-native consonant clusters. Following

Peperkamp and Dupoux (2003); Peperkamp (2007); Wilson et al. (2014); Zhao and Berent (2016), we investigated the role of phonetic details, particularly, stop burst duration and intensity.

5.2 The role of phonetic details: stop burst duration and intensity

The results from both perception and production experiments appear to be consistent with the second hypothesis: if Mandarin speakers are sensitive to phonetic details, we expect that they will also perceive and produce non-native CCs with other types of modifications. We tested two acoustic details in the current study: stop burst duration and intensity. Results show that when the burst intensity was lower, Mandarin speakers were more likely to misperceive the clusters. In the ABX discrimination experiment, Mandarin speakers were highly sensitive to the difference between the medial-/pt/ cluster and its corresponding control medial-/pət/, even significantly more so than for the cluster /nk/, which exists in Mandarin. We measured burst intensity and found that /p/ does indeed have the lowest intensity burst of all the stops, and a further analysis indicated that there is a significant effect of burst intensity on the discrimination between non-native CCs and CVC sequences.

However, successful discrimination does not show that Mandarin speakers perceived CCs exactly as in the target language. The follow-up transcription experiment contributed additional information about perception. We observed that the medial-/pt/ cluster was transcribed with feature change and deletion 46.3% of the time. In the same cluster, in the prompted production task, [p] was changed or deleted 65% of the time. Stop consonants are known to differ in the energy of the released airflow depending on their place of articulation (Stevens, 1998). The

energy of the release burst for a voiceless bilabial plosive /p/ is weaker than it is for velar /k/ and alveolar /t/. The weak release burst of /p/ observed in our data is therefore unsurprising, and it is partially consistent with the results of Wilson et al. (2014), who found that lower burst intensity and shorter stop burst duration resulted in consonant deletion and change. In our study, Mandarin participants were sensitive to burst intensity, but not to burst duration.

5.3 Correct production of non-native CCs

An important finding in the current study is that monolingual speakers of Mandarin correctly produced the non-native consonant clusters 13% of the time (rates ranging from approximately 1% to over 29% across cluster types). Correct production suggests that Mandarin speakers may have perceived non-native consonant clusters without any epenthetic vowels between the two consonants or other modifications. Monolingual speakers, who had never heard the consonant clusters before, faithfully decoded the phonetic details in the Russian stimuli by relying on the auditory inputs. Unlike in the case of Japanese speakers, who, as reported in Dupoux et al. (1999), consistently perceived an illusory vowel in non-native consonant clusters, the Mandarin speakers in our study were less consistently 'deaf' to the non-native clusters. They sometimes perceived them correctly and produced them without a vocoid.

Since our study is based on acoustic data, our hypotheses, predictions, and interpretation of the results can only refer to the presence or absence of a vocoid in the cluster. We cannot draw any conclusions regarding the details of the articulatory production. We can, however, discuss the possible articulatory scenarios that can arise in correct acoustic production.

There are two possibilities for explaining correct production. First, even when

no vocoid is observed between the consonants, a vowel gesture may still be produced. In Mandarin, as discussed in 1.4, vowels can be devoiced after a voiceless/aspirated onset, in unstressed position, and in the context of a low tone. These conditions correspond to the phonetic context of the consonant clusters tested here, where C1 is voiceless, and the clusters are in unstressed position, either preceding or following a stressed syllable. If Mandarin speakers correctly perceive the non-native consonant clusters, they should tend to make their production as close as possible to what they have heard. But since Mandarin speakers have only learned to produce a CVC timing pattern, they may continue to produce that same pattern in response to clusters, possibly with a suppression of voicing. Under this scenario, they would be adapting to cluster production by producing a devoiced vowel. The second possibility for the correct articulation of a cluster is that monolingual speakers perceive the non-native consonant clusters correctly, and also produce them correctly, without a vowel gesture between the consonants. Deciding between these two interpretations will only be possible in a future, planned study involving articulatory data.

5.4 Gestural coordination patterns in the production of non-native CCs

We mentioned that an important question in our study is determining the gestural coordination patterns that Mandarin speakers employ in producing consonant clusters. This question is only in part answered in our study, by examining acoustic timing patterns. In our study, we observed three types of acoustic productions of non-native consonant clusters: with a vocoid, with voicing, and without vocoid (correct production).

Production with a vocoid may result from two possible articulatory events,

as discussed in Section 1.3.2. One is the production of a vowel that would be identical to a lexical vowel. This is what we consider an epenthetic vowel. The other possibility is that the vocoid observed in the acoustic signal is due to an 'open transition', resulting from minimally overlapping consonantal constrictions. However, we argued, based on our results, that the vowels between two consonants are lexical vowels, not 'transitions'. Their acoustic properties are similar to a reduced vowel in Mandarin. The acoustic measurements we used, based on Davidson (2006a), revealed that the duration of the vocoids was significantly shorter than the duration of vowels in the control CVC sequences in pre-stress syllables, but in post-stress syllables the durations were comparable. Moreover, the F1 and F2 values of the vocoids indicate a schwa-like quality (mean F1 = 476 Hz, F2 = 1628 Hz). The vocalic space of the vowels 'covers' that of the vocoids. Only the F1 of the inserted vocoids in CCs was lower than that of the vowels in CVCs. Moreover, by measuring the entire duration of the CC interval, we found that Mandarin speakers used similar CC timing patterns in the production of CCs with a vocoid and their CVC controls, except for the cluster /tp/. This suggests that they may have used, in both conditions, the only consonant-to-consonant timing known to them, namely the one which they use in native CVC sequences.

The analysis of acoustic timing lags suggests that Mandarin speakers maintained the native CVC timing pattern in the production of non-native CCs with vocoids. Even though they were sometimes able to suppress the vocoid acoustically, they did not seem to adjust the articulatory timing. Instead, they retained the timing pattern by increasing the release duration of the preceding consonant. This finding is not consistent with that of Davidson (2006a) for English speakers. She claimed that the presence of vocoids in English speakers' production is not due to lack of motor skill with non-native consonant clusters, but rather to the speaker/listeners' successful adaptation of their native timing pattern to that of

non-native consonant clusters. Unlike English speakers, Mandarin speakers do not have any native timing pattern associated with consonant clusters. Thus, we do not necessarily expect that the vocoids in their production of non-native CC should always be 'transitions'. We saw, in fact, that Mandarin speakers interpret certain properties of C1 release as a vowel. We thus expect that Mandarin speakers may produce a reduced vowel, not necessarily a full lexical vowel, when repeating the sequences they hear.

In Mandarin, reduced vowels always occur in the context of a neutral tone in unstressed position. Lin and Yan (1980) found that reduced vowels in Mandarin occur in toneless and unstressed syllables, and have a duration 50% shorter than that of lexical vowels. Reduced vowels tend to be centralized, resembling a schwa [ə]. According to Duanmu (2007), a high vowel with a low or neutral tone can be devoiced after an aspirated stop. In producing non-native consonant clusters, Mandarin speakers may therefore employ the native coordination patterns they have learned for reduced vowels. They may decrease the magnitude or duration of the vocalic gesture, and they may increase the overlap between the vocalic gesture and the gestures of the flanking consonants. However, the vowel gesture may still be present, even though they were sometimes able to suppress the vocoid acoustically. They may not have adjusted the articulatory timing enough to match the timing of a CC cluster.

The remaining two types of acoustic productions observed involve production without a vocoid, and production with voicing only. Unlike in the case of production with a vocoid, Mandarin speakers did not use native CVC timing patterns for these two realizations. These two types of production had more variations than production with a vocoid. As discussed earlier, we cannot decide, based on our acoustic data, whether the 'correct' production without a vocoid is due to vowel devoicing or to the actual faithful production of consonant clusters. The corre-

sponding gestural coordination patterns can only be examined with articulatory data. Overall, our findings suggest considerably high variability in gestural coordination across different types of production, showing that Mandarin speakers used various strategies to produce non-native consonant clusters.

5.5 Conclusion and future study

This dissertation investigated the following question: how do native speakers of languages with simple or moderately complex syllable structure perceive and produce non-native clusters (CCs)? by testing monolingual speakers of Mandarin Chinese. We found that Mandarin speakers sometimes perceived and produced an epenthetic vowel (or vocoid [2]) between two consonants. This vocoid in the production of non-native CCs is similar to a reduced vowel in Mandarin, with shorter duration and the vowel space centralized to a schwa-like quality. In addition, the duration of the CC interval with a vocoid is similar to the duration of the CVC interval with a vowel. Interestingly, Mandarin speakers sometimes produced non-native CC sequences without vocoids or any other modifications, and sometimes they produced non-native CCs with only a period of voicing, only relying on auditory inputs. These findings suggest that Mandarin speakers are highly sensitive to the phonetic details in the inputs, and try to decode the phonetic details as faithfully as possible. Correct production indicates that they may successfully decode the phonetic details in the input and perceive the non-native CCs correctly, without a vocoid or any other modifications. Furthermore, these different types of productions led to variability in gestural coordination patterns: Mandarin speakers

[2]Recall that in production, these acoustic vowels have received two different treatments. Under one view they are considered epenthetic vowels; under an other view, they are considered acoustic transitions between consonants. The two types of vowels may even be mixed in the production of non-native CCs (Shaw and Davidson, 2011). Thus, we use a phonetic terminology 'vocoid' to refer to such vowels in the case of production of non-native consonant clusters.

maintained the native CVC timing pattern in production with vocoids, however, in the production of CCs without vocoids and with voicing, they used a different strategy from the native one. It should be noted that these conclusions are drawn only from acoustic data. The corresponding gestural coordination patterns should be further examined with articulatory data. To this end, we collected electromagnetic articulography (EMA) data from two speakers of Mandarin, and are planning a pilot study based on these data, involving the same stimuli from the acoustic study. Both speakers have also participated in the experiments presented here, which will allow for an informative comparison of their respective perception and production patterns. In the prompted production task, one of the participants produced a vocoid between two consonants more than 60% of the time, while the other produced such a vocoid less than 40% of the time. Specifically, the addition of articulatory data will help to understand how the gestural coordination patterns relate to the presence or absence of a vocoid in the acoustic signal.

Appendices

Appendix A

Questionnaire

A.1 Linguistic questionnaire

Linguistic Questionnaire
调查问卷

Name : 姓名	
Age: 年龄	
Gender : 性别	
Education: 文化程度	
City 来自哪个地区 / 城市 / 省？	
Which language did you speak growing up ? Do you speak any dialect ? Which one ? 母语是什么？还说其他方言吗？什么方言？	
Have you ever lived in other countries ? If yes, which ones ? When ? 有没有在其他国家居住过？如果有，哪个国家？什么时候？	
What other languages have you studied/ do you speak ? 有没有学习过其他语言？	
If so, which level ? 如果有，什么水平？	
The number of years you have studied other languages 学习多久了？	
Do you have any speech or hearing problems? 有没有发音或者听力问题？	

Appendix B

Stimuli measurements

B.1 ABX discrimination experiment

Position	Cluster	Nativeness	Segmental durations								Intensity	Vowel formants					
			Total	V	Burst	IBI	IPI	CC	V1	V3	C1 Intensity	F1 (V)	F2 (V)	F1 (V1)	F2 (V1)	F1 (V3)	F2 (V3)
initial	kt	CVC	490	53	29	217	82	NA	124	50	79	741	1616	927	1642	676	1215
			522	64	31	218	95	NA	137	58	82	789	1684	942	1650	590	1024
		CC	472	NA	42	172	42	NA	128	40	76	NA	NA	827	1597	620	1220
			468	NA	48	158	48	NA	137	54	74	NA	NA	964	1435	562	1143
	pt	CVC	435	56	8	150	64	NA	112	54	70	848	1603	855	1614	706	1318
			431	58	13	162	71	NA	109	34	70	840	1529	854	1630	704	1408
		CC	405	NA	40	111	40	NA	129	37	72	NA	NA	821	1634	536	1202
			391	NA	42	118	42	NA	120	34	72	NA	NA	858	1446	530	1277
	tk	CVC	458	58	28	185	86	NA	114	48	77	700	1583	853	1643	570	1072
			470	64	33	196	97	NA	112	44	77	772	1608	923	1545	663	1226
		CC	433	NA	61	139	61	NA	117	60	78	NA	NA	792	1787	600	1275
			423	NA	75	130	77	NA	126	52	75	NA	NA	852	1519	561	1094
	tp	CVC	447	55	18	187	73	NA	125	25	77	734	1600	861	1523	600	1200
			450	60	22	191	82	NA	107	47	77	768	1405	820	1444	643	1358
		CC	466	NA	50	153	50	NA	154	45	78	NA	NA	827	1404	668	1312
			478	NA	54	166	54	NA	151	46	75	NA	NA	862	1371	660	1383

Position	Cluster	Structure	Segmental durations								Intensity	Vowel formants					
			Total	Reduced V	C1 Burst	IBI	IPI	CC	V1	V3	C1 Intensity	F1 (reduced V)	F2 (reduced V)	F1 (V1)	F2 (V1)	F1 (V3)	F2 (V3)
medial	kt	CVC	382	51	18	122	69	212	143	28	82	681	1661	926	1524	575	1570
			409	53	22	142	75	236	143	30	82	595	1669	856	1413	697	1541
		CC	346	NA	36	91	36	179	128	39	81	NA	NA	967	1461	732	1452
			340	NA	38	88	38	179	116	46	83	NA	NA	824	1553	690	1426
	pt	CVC	419	72	18	139	90	233	148	38	80	741	1520	973	1496	565	1507
			410	69	20	151	88	237	138	35	80	722	1544	964	1326	605	1505
		CC	331	NA	33	65	33	173	113	44	73	NA	NA	884	1469	585	1695
			347	NA	33	77	33	179	130	37	72	NA	NA	941	1434	612	1524
	tk	CVC	415	57	24	156	82	238	150	28	80	655	1663	937	1565	559	1229
			401	59	25	139	84	211	154	36	78	718	1678	869	1499	562	1272
		CC	315	NA	42	93	42	163	116	36	77	NA	NA	869	1495	574	1292
			368	NA	43	107	43	187	139	42	77	NA	NA	869	1471	520	1281
	tp	CVC	416	48	17	152	66	237	138	42	81	714	1726	777	1462	564	1172
			410	46	20	155	67	220	152	38	79	746	1625	920	1419	563	1208
		CC	355	NA	42	106	42	171	133	51	74	NA	NA	857	1427	630	1224
			336	NA	47	127	47	179	115	42	77	NA	NA	895	1527	650	1223
	nk	CVC	420	80	NA	NA	NA	240	144	36	NA	737	1696	673	1511	548	1239
			420	90	NA	NA	NA	230	154	36	NA	757	1763	819	1511	560	1200
		CC	366	NA	NA	NA	NA	197	129	40	NA	NA	NA	767	1291	661	1261
			383	NA	NA	NA	NA	196	134	53	NA	NA	NA	869	1513	712	1423

B.2 Repetition and transcription experiments

C1/C2 manner	Position	Cluster	Nativeness	Total	Burst	IBI	IPI	Intensity	V
SS	medial	kt	CVC	397	31	140	69	64	37
				383	28	134	64	63	36
			CC	323	30	88	30	56	NA
				322	37	87	37	55	NA
		pt	CVC	374	16	137	74	74	58
				368	18	154	74	70	56
			CC	310	18	74	18	50	NA
				300	13	82	13	44	NA
		tk	CVC	380	30	132	59	54	29
				404	21	135	56	62	35
			CC	349	22	89	22	48	NA
				324	16	89	16	57	NA
		tp	CVC	365	19	121	55	61	36
				379	16	141	63	69	46
			CC	332	11	73	11	47	NA
				334	12	75	10	44	NA

C1/C2 manner	Position	Cluster	Nativeness	Total	Burst	IBI	IPI	Intensity	V
SN	initial	kt	CVC	424	25	190	84	70	59
				459	33	193	100	64	66
			CC	379	29	118	29	56	NA
				402	37	110	37	51	NA
		pt	CVC	427	13	178	92	68	79
				433	14	180	96	65	82
			CC	372	22	89	22	44	NA
				396	22	118	22	42	NA
		tk	CVC	442	19	175	105	62	86
				461	13	173	107	68	95
			CC	364	30	93	30	57	NA
				387	34	97	34	58	NA
	medial	kn	CVC	442	35	NA	87	68	52
				399	41	NA	83	64	42
			CC	356	71	NA	71	43	NA
				370	90	NA	90	45	NA
	initial	kn	CVC	422	39	NA	97	61	58
				445	38	NA	111	64	72
			CC	394	74	NA	74	43	NA
				404	83	NA	83	45	NA

C1/C2 manner	Position	Cluster	Nativeness	Total	Burst	IBI	IPI	Intensity	V
SL	medial	kl	CVC	418	51	NA	77	62	26
				404	44	NA	87	56	43
			CC	358	95	NA	95	39	NA
				348	98	NA	98	43	NA
		pl	CVC	343	20	NA	63	65	43
				358	22	NA	74	58	51
			CC	286	14	NA	14	59	NA
				300	39	NA	39	53	NA
	initial	kl	CVC	420	34	NA	111	63	77
				436	44	NA	112	59	68
			CC	403	65	NA	65	57	NA
				409	70	NA	70	61	NA
		pl	CVC	430	15	NA	103	66	88
				410	15	NA	99	65	84
			CC	353	35	NA	35	53	NA
				384	29	NA	29	56	NA

C1/C2 manner	Position	Cluster	Nativeness	Total	Burst	IBI	IPI	Intensity	V
LS	medial	lk	CVC	383	NA	NA	113	63	56
				404	NA	NA	132	61	56
			CC	385	NA	NA	59	62	NA
				341	NA	NA	63	58	NA
		lp	CVC	392	NA	NA	88	54	36
				402	NA	NA	108	57	60
			CC	331	NA	NA	68	69	NA
				319	NA	NA	63	61	NA
NS (native)	medial	nk	CVC	377	NA	NA	121	57	60
				345	NA	NA	134	65	70
			CC	324	NA	NA	97	60	NA
				292	NA	NA	79	54	NA

Bibliography

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52:317–332.

Bao, Z. (1990). Fanqie languages and reduplication. *Linguistic Inquiry*, 21(3):317–350.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*.

Bellik, J. (2018). An acoustic study of vowel intrusion in Turkish onset clusters. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 9(1).

Berent, I., Lennertz, T., and Balaban, E. (2012a). Language universals and misidentification: A two-way street. *Language and Speech*, 55(3):311–330.

Berent, I., Lennertz, T., Jun, J., Moreno, M. A., and Smolensky, P. (2008). Language universals in human brains. *Proceedings of the National Academy of Sciences*, 105(14):5321–5325.

Berent, I., Lennertz, T., and Rosselli, M. (2012b). Universal phonological restrictions and language-specific repairs: Evidence from spanish. *The Mental Lexicon*, 13:275–305.

Berent, I., Lennertz, T., Smolensky, P., and Vaknin-Nusbaum, V. (2009). Listeners' knowledge of phonological universals: Evidence from nasal clusters. *Phonology*, 26(1):75–108.

Berent, I., Steriade, D., Lennertz, T., and Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, 104(3):591–630.

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

Boersma, P. and Weenink, D. (2014). *Praat: Doing phonetics by computer (version 5. 3. 84)*. http://www.praat.org/.

Broselow, E. (1983). *Nonobvious transfer: On predicting epenthesis errors*. Interlanguage Phonology: The Acquisition of a Second Language Sound System. Cambridge, MA: Newbury House.

Broselow, E. (2015). The typology of position-quality interactions in loanword vowel insertion. *Capturing Phonological Shades*, pages 292–319.

Broselow, E., Chen, S.-I., and Wang, C. (1998). The emergence of the unmarked in second language phonology. *Studies in second language acquisition*, 20(2):261–280.

Broselow, E. and Finer, D. (1991). Parameter setting in second language phonology and syntax. *Second Language Research*, 7(1):35–59.

Browman, C. P. and Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. *Papers in laboratory phonology I: Between the grammar and physics of speech*, pages 341–376.

Browman, C. P. and Goldstein, L. (1992a). Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180.

Browman, C. P. and Goldstein, L. (1992b). "targetless" schwa: an articulatory analysis. *Papers in laboratory phonology II: Gesture, segment, prosody*, pages 26–56.

Browman, C. P. and Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252.

Catford, J. C. (1985). 'rest' and 'open transition' in a systemic phonology of English. *Systemic perspectives on discourse*, 1:333–348.

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

Cheng, C.-C. (1973). *A synchronic phonology of Mandarin Chinese*, volume 4. Walter de Gruyter.

Childers, D. G. (1978). *Modern spectrum analysis*. IEEE Computer Society Press.

Chitoran, I. (1999). Accounting for sonority violations: the case of Georgian consonant sequencing. In *Proceedings of the 14th International Congress of Phonetic Sciences. Berkeley: Department of Linguistics, University of California, Berkeley*, pages 101–104.

Chitoran, I., Goldstein, L., and Byrd, D. (2002). Gestural overlap and recoverability: Articulatory evidence from Georgian. In Carlos, G. and Natasha, W., editors, *Laboratory Phonology 7*, pages 419–447. Mouton de Gruyter.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. *Papers in laboratory phonology*, 1:283–333.

Dancey, C. and Reidy, J. (2008). *Statistics Without Maths for Psychology: Using Spss for Windows*. Prentice Hall International (UK) Ltd., Hertfordshire, UK, UK, 4th edition.

Davidson, L. (2005). Addressing phonological questions with ultrasound. *Clinical Linguistics & Phonetics*, 19(6-7):619–633.

Davidson, L. (2006a). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics*, 34(1):104–137.

Davidson, L. (2006b). Phonotactics and articulatory coordination interact in phonology: Evidence from nonnative production. *Cognitive Science*, 30(5):837–862.

Davidson, L. (2007). The relationship between the perception of non-native phonotactics and loanword adaptation. *Phonology*, 24(2):261–286.

Davidson, L., Jusczyk, P., and Smolensky, P. (2004). The initial and final states: Theoretical implications and experimental explorations of richness of the base. *Constraints in phonological acquisition*, pages 321–368.

Davidson, L., Martin, S., and Wilson, C. (2015). Stabilizing the production of non-native consonant clusters with acoustic variability. *The Journal of the Acoustical Society of America*, 137(2):856–872.

Davidson, L. and Shaw, J. A. (2012). Sources of illusion in consonant cluster perception. *Journal of Phonetics*, 40(2):234–248.

de Jong, K. and Park, H. (2012). Vowel epenthesis and segment identity in Korean learners of English. *Studies in Second Language Acquisition*, 34(1):127–155.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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

Duanmu, S. (2009). *Syllable structure: The limits of variation*. Oxford University Press.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., and Mehler, J. (1999). Epenthetic vowels in japanese: A perceptual illusion? *Journal of experimental psychology: human perception and performance*, 106:1568–1578.

Durvasula, K., Huang, H.-H., Uehara, S., Luo, Q., and Lin, Y.-H. (2018). Phonology modulates the illusory vowels in perceptual illusions: Evidence from Mandarin and English. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 9(1).

Durvasula, K. and Kahng, J. (2015). Illusory vowels in perceptual epenthesis: the role of phonological alternations. *Phonology*, 32(3):385–416.

Easterday, S. (2017). *Highly complex syllable structure: a typological study of its phonological characteristics and diachronic development*. PhD thesis, University of New Mexico.

Fan, Y. (2011). Articulatory timing of english consonant clusters in the coda positions: a study of Chinese-English interlanguage. Master's thesis, University of Victoria.

Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 4:22–30.

Fant, G. (1975). Non-uniform vowel normalization. *STL-QPSR*, 16(2-3):1–19.

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In Strange, W., editor, *Speech perception and linguistic experience: issues in cross-language research*, pages 233–273. Timonium, MD: York Press.

Gafos, A. I. (2002). A grammar of gestural coordination. *Natural Language & Linguistic Theory*, 20(2):269–337.

Gafos, A. I., Hoole, P., Roon, K., and Zeroual, C. (2010). Variation in overlap and phonological grammar in Moroccan Arabic clusters. In Fougeron, C., Kühnert, B., D'Imperio, M., and Vallée, N., editors, *Laboratory Phonology 10: Variation, Detail and Representation*, pages 657–698. Mouton de Gruyter.

Goldstein, L., Nam, H., Saltzman, E., and Chitoran, I. (2009). Coupled oscillator planning model of speech timing and syllable structure. In Fant, G., Fujisaki, H., and Shen, J., editors, *Frontiers in Phonetics and Speech Science. Beijing*, pages 239–250. The Commercial Press Beijing.

Green, D. M. and Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Greenberg, J. H. (1965). Some generalizations concerning initial and final consonant sequences. *Linguistics*, 3(18):5–34.

Guo, X. and Nogita, A. (2013). Lexical schwa and inserted schwa produced by Mandarin Chinese EAL learners. *Working Papers of the Linguistics Circle*, 23(1):81–109.

Hadfield, J. D. et al. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22.

Hall, N. (2006). Cross-linguistic patterns of vowel intrusion. *Phonology*, 23(3):387–429.

Hallé, P., Segui, J., Dominguez, A., and Cuetos, F. (2014). Special is especial but stuto is not astuto: Perception of prothetic/e/in speech and print by speakers of Spanish. *Psicolinguıstica en Espanol. Homenaje a Juan Seguı*, pages 31–47.

Hallé, P. A., Segui, J., Frauenfelder, U., and Meunier, C. (1998). Processing of illegal consonant clusters: A case of perceptual assimilation? *Journal of experimental psychology: Human perception and performance*, 24(2):592.

Hamilton, W. S. (1980). *Introduction to Russian phonology and word structure*. Slavica Pub.

Hancin-Bhatt, B. and Bhatt, R. M. (1997). Optimal l2 syllables: Interactions of transfer and developmental effects. *Studies in second language acquisition*, 19(3):331–378.

Hansen, J. G. (2001). Linguistic constraints on the acquisition of English syllable codas by native speakers of Mandarin Chinese. *Applied Linguistics*, 22(3):338–365.

Hautus, M. J. and Meng, X. (2002). Decision strategies in the ABX (matching-to-sample) psychophysical task. *Perception & psychophysics*, 64(1):89–106.

Hirose, H. (1971). The activity of the adductor laryngeal muscles in respect to vowel devoicing in Japanese. *Phonetica*, 23(3):156–170.

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

Jun, S.-A., Beckman, M., Niimi, S., and Tiede, M. (1997). Electromyographic evidence for a gestural-overlap analysis of vowel devoicing in Korean. *The Journal of Speech Sciences (The Korean Society of Speech Sciences)*, 1:153–200.

Kadlec, H. (1999). Statistical properties of d'and β estimates of signal detection theory. *Psychological Methods*, 4(1):22.

Kochetov, A. (2006). Syllable position effects and gestural organization: Articulatory evidence from Russian. *Papers in laboratory phonology*, 8:565–588.

Kochetov, A., Pouplier, M., and Son, M. (2007). Cross-language differences in overlap and assimilation patterns in Korean and Russian. In *Proceedings of the XVI International Congress of Phonetic Sciences, Saarbrücken*, pages 1361–1364.

Krom, Guus, D. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech, Language, and Hearing Research*, 36(2):254–266.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13).

Lee, W.-S. and Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association*, 33(1):109–112.

Lee-Kim, S.-I. (2014). Revisiting Mandarin 'apical vowels' : An articulatory and acoustic study. *Journal of the International Phonetic Association*, 44(3):261–282.

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

Lentz, T. O. and Kager, R. W. (2015). Categorical phonotactic knowledge filters second language input, but probabilistic phonotactic knowledge can still be acquired. *Language and speech*, 58(3):387–413.

Lin, M. and Yan, J. (1980). Beijinghua qingsheng de shengxue xingzhi [acoustic characteristics of neutral tone in Beijing Mandarin]. *Dialect*, 3:166–178.

Lin, Y.-H. (1989). *Autosegmental treatment of segmental processes in Chinese phonology*. PhD thesis, University of Texas at Austin.

Lin, Y.-H. (2007). *The Sounds of Chinese*, volume 1. Cambridge University Press.

Luo, S. (2017). Gestural overlap across word boundaries: Evidence from English and Mandarin speakers. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 62(1):56–83.

Macmillan, N. A. (2002). Signal detection theory. *Stevens' handbook of experimental psychology*, 4:43–90.

Macmillan, N. A. and Creelman, C. D. (2005). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maddieson, I. (1984). *Patterns of sounds*. Cambridge University Press.

Maddieson, I. (2013). Syllable structure. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Miao, R. (2005). *Loanword adaptation in Mandarin Chinese: Perceptual, phonological and sociolinguistic factors*. PhD thesis, Stony Brook University.

Moreton, E. (2002). Structural constraints in the perception of English stop-sonorant clusters. *Cognition*, 84(1):55–71.

Myers, L. and Sirois, M. J. (2006). Spearman correlation coefficients, Differences between. *Encyclopedia of statistical sciences*.

Öhman, S. E. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39(1):151–168.

Padgett, J. and Tabain, M. (2005). Adaptive dispersion theory and phonological vowel reduction in Russian. *Phonetica*, 62(1):14–54.

Peperkamp, S. (2007). Do we have innate knowledge about phonological markedness? Comments on Berent, Steriade, Lennertz, and Vaknin. *Cognition*, 104(3):631–637.

Peperkamp, S. and Dupoux, E. (2003). Reinterpreting loanword adaptations: the role of perception. In Solé, M. J., Recasens, D., and Romero, J., editors, *Proceedings of the 15th international congress of phonetic sciences*, volume 367, page 370, Barcelona: Universitat Autónoma de Barcelona.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, 24(2):175–184.

Pike, K. L. (1943). *Phonetics: a critical analysis of phonetic theory and a technique for the practical description of sounds*. University of Michigan Press.

Pitt, M. A. (1998). Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception & psychophysics*, 60(6):941–951.

Polivanov, E. (1931). La perception des sons d' une langue étrangère. *Travaux du Cercle linguistique de Prague*, 4:79–96.

Pouplier, M. and Goldstein, L. (2005). Asymmetries in the perception of speech production errors. *Journal of Phonetics*, 33(1):47–75.

Press, W. H., Teukolsky, S. A., Vetterling, W., and Flannery, B. (1992). Numerical recipes in C: the art of scientific computing. *Cambridge University Press*.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Ridouane, R. and Fougeron, C. (2011). Schwa elements in Tashlhiyt word-initial clusters. *Laboratory Phonology*, 2(2):275–300.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Selkirk, E. (1984). On the major class features and syllable theory. *Language sound structure*.

Shaw, J. A. and Davidson, L. (2011). Perceptual similarity in input–output mappings: A computational/experimental study of non-native speech production. *Lingua*, 121(8):1344–1358.

Shaw, J. A. and Kawahara, S. (2018). The lingual articulation of devoiced/u/in Tokyo Japanese. *Journal of Phonetics*, 66:100–119.

Silverman, D. (1992). Multiple scansions in loanword phonology: evidence from Cantonese. *Phonology*, 9(2):289–328.

Smolensky, P. (2006). Optimality in phonology ii: Markedness, feature domains, and local constraint conjunction. *The harmonic mind: From neural computation to optimality-theoretic grammar*, 2:27–160.

Sproat, R. and Fujimura, O. (1993). Allophonic variation in English/l/and its implications for phonetic implementation. *Journal of phonetics*, 21(3):291–311.

Stevens, K. N. (1998). *Acoustic phonetics*, volume 30. MIT press.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88(1):97–100.

Wang, C. (1995). *The acquisition of English word-final obstruents by Chinese speakers*. Phd thesis, State University of New York at Stony Brook.

Weinberger, S. H. (1997). Minimal segments in second language phonology. *Second language speech: Structure and process*, pages 263–312.

Wilson, C., Davidson, L., and Martin, S. (2014). Effects of acoustic–phonetic detail on cross-language speech production. *Journal of Memory and Language*, 77:1–24.

Wright, R. A. (1996). *Consonant clusters and cue preservation in Tsou*. PhD thesis, Department of Linguistics, University of California, Los Angeles.

Yanagawa, M. (2006). *Articulatory timing in first and second language: A cross-linguistic study*. PhD thesis, Yale University.

Yip, M. (1993). Cantonese loanword phonology and Optimality Theory. *Journal of East Asian Linguistics*, 2(3):261–291.

Yip, M. (2003). Casting doubt on the onset–rime distinction. *Lingua*, 113(8):779–816.

Zhao, X. and Berent, I. (2016). Universal restrictions on syllable structure: Evidence from Mandarin Chinese. *Journal of psycholinguistic research*, 45(4):795–811.

Zsiga, E. C. (2000). Phonetic alignment constraints: Consonant overlap and palatalization in English and Russian. *Journal of phonetics*, 28(1):69–102.

Zsiga, E. C. (2003). Articulatory timing in a second language: Evidence from Russian and English. *Studies in Second Language Acquisition*, 25(3):399–432.