The Effect of Vowel Duration on Native Mandarin Listeners' Perception of Australian English Vowel Contrasts in Voiced and Voiceless Coda Contexts

Shuting Liu

B.A., China University of Political Science and Law

Department of Linguistics

Faculty of Human Sciences

Macquarie University

A thesis submitted in fulfilment of the requirement for the degree of Master of Research (Linguistics)

10th October, 2016

Table of Contents

Abstract

English vowels differ in spectral and durational properties. Mandarin learners of English rely mainly on duration to categorize some English vowel contrasts (e.g., /i/-/ɪ/), whereas native listeners predominantly use spectral cues. As Mandarin is a language that does not have vowel duration contrast, Mandarin listeners' use of duration cues could be attributed to the development of duration categories during the L2 learning process. To examine whether Mandarin learners of English can develop perceptual categories based on duration only, this study investigates their perception of Australian English (AusE) vowel contrasts /ɐ/-/ɐː/ and /ɔ/-/oː/ in both /hVd/ and /hVt/ context. AusE vowels /ɐ/ and /ɐː/ are mainly contrasted in length, and /ɔ/ and /oː/ are contrasted in both length and spectrum. All vowels are longer before a voiced than a voiceless coda consonant. Duration of the four vowels in /hVd/ and /hVt/ context is varied in 11 steps with endpoints 85ms and 335ms and the interval between steps 25ms. With 4 vowels, 2 contexts, and 11 steps, this yields 88 stimuli. Participants are Mandarin listeners who have studied in Australia for more than 6 months and native AusE listeners. Participants are asked to perform a 2-alternative forced choice perceptual categorization task for each contrast in each condition. The findings are: 1) both groups show categorical perception of duration along the /ɐ/ and /ɐː/ continua, but the category boundary differs between groups in location and steepness; 2) Mandarin listeners are still influenced by duration perceiving /ɔ/ and /oː/ continua, whereas AusE listeners could categorize the stimuli based on spectral features; 3) coda voicing can influence both groups' category boundary. The results from this study will shed further light on the degree to which Mandarin listeners can and will use vowel duration in the acquisition of English vowels.

Declaration

I hereby declare that this thesis has not been submitted for a higher degree to any other university or institution. I have made every effort to clearly indicate the sources of information used and acknowledge the extent to which the work of others has been used in the text. The research presented in this thesis has been approved by the Macquarie University Faculty of Human Sciences Research Ethics Sub-Committee (Ref: 5201600131).

Shuting Liu

10.10.2016

Chapter 1 Introduction

English vowels differ in spectral and durational properties. Native English listeners predominantly use spectral cues to perceive their native vowels. Second-language (L2) learners of English, on the contrary, primarily adopt the duration cue to categorize some English vowel contrasts (e.g., /i/-/ɪ/). Several explanations have been proposed to account for this duration reliance. First language (L1) transfer theories suggest that L2 listeners have directly transferred their L1 experience with duration to L2 perception. The Desensitisation Hypothesis indicates that duration is a highly salient cue employed by default when spectral differences are insufficient for vowel identification. A developmental approach argues that L2 learners from an L1 without phonological vowel duration rely on duration because they form duration-based categories (i.e. long and short) along with their exposure to the L2.

The current study aims to specifically explore the developmental approach by examining Mandarin listeners' perception of Australian English (AusE) vowel contrasts /ɐ/-/ɐː/ and /ɔ/-/oː/ in both /hVd/ and /hVt/ context. Questions of interest are 1) whether Mandarin learners of English can and will develop perceptual categories based on duration and 2) to which degree Mandarin listeners will be influenced by duration cue in perceiving L2 vowels.

Mandarin is a language that does not have vowel duration contrast. AusE is an English dialect that has phonemic vowel length. AusE vowels /ɐ/ and /ɐː/ are mainly contrasted in duration. So if listeners have perceptual categories based on duration, the category boundary should be clearly detected during the perception of /ɐ/-/ɐː/ duration continua by listeners of both L1 and L2 English. By contrast, AusE vowels /ɔ/ and /oː/ differ in both length and spectrum. As they fall outside any vowel category of Mandarin, the spectral difference might be sufficient for L1 English listeners to differentiate between the vowels but insufficient for Mandarin listeners to identify this contrast. Thus Mandarin listeners' performance on /ɔ/-/oː/ duration continua could show how heavily duration affects their perception of this contrast.

Although the inherent duration of an AusE vowel is either short or long, the vowel is longer before a voiced than a voiceless coda consonant. This difference may result in a shift of the duration perceptual boundary. Thus the present study investigates the vowels in both voiced and voiceless contexts.

Chapter 2 Literature Review

This chapter presents theories and issues that are central to this thesis through a review of how second-language (L2) learners perceive L2 sounds and how vowel duration affects the perception. Section 2.1 provides a summary of L2 sounds perception theories including three prominent speech models in L2 research: the Speech Learning Model (SLM; Flege, 1995), the Perceptual Assimilation Model's L2 extension (PAM-L2; Best & Tyler, 2007), and the Second Language Linguistic Perception model (L2LP; Escudero, 2005). Those models account for how L2 learners acquire L2 sounds in general. This is followed in Section 2.2 by a review of studies concerning different strategies employed by L1 and L2 listeners to identify certain English vowel contrasts. L1 listeners (i.e. native English listeners) may rely more on vowel quality rather than quantity to make certain vowel distinction, whereas L2 listeners tend to rely more on vowel duration for these contrasts irrespective of their various native language background. Hypotheses for why vowel duration has such an effect on L2 perception are presented at the end of this section. Section 2.3 argues that these hypotheses can be tested by examining Mandarin L2 listeners' perception of Australian English vowels. In this section we compare the vowel system of Mandarin Chinese with that of Australian English and apply the L2LP to make predictions about Mandarin listeners' perception of Australian English vowels. The review formulates the specific research questions of the present study in Section 2.4.

2.1 Theories for Second-language Speech Perception

2.1.1 Overview

Second-language (L2) learners are reported to have difficulties identifying and discriminating some pairs of sounds in the L2. A typical example is the perception of American English /r/-/l/ contrast by adult Japanese learners of English. Japanese listeners who start to learn English after childhood often fail to differentiate English syllables starting with /r/ from those with /l/, such as "right" from "light" (e.g., Cochrane, 1980; Goto, 1971). The discrimination performance improves when Japanese learners' experience with English increases (Aoyama, Flege, Guion, Akahane-Yamada, & Yamada, 2004; MacKain, Best, & Strange, 1981), but it remains a question whether they are able to discriminate the sound pair as well as native English listeners do (MacKain et al., 1981).

The problems listeners have with L2 sounds are generally attributed to their prior linguistic experience, i.e. the first language (L1) background. For instance, according to

Trubetzkoy (1939/1969), listeners perceive L2 sounds through their L1 phonological system and hence tend to associate L2 sounds with the L1 sounds (see also Polivanov, 1931 as cited in Escudero, 2005). The association results in difficulties in discriminating L2 sounds or in perceiving L2 sounds in a native-like way. More recent studies have attributed L2 listeners' perceptual difficulties to the perceptual system's insensitivity to L2 sounds because of L1 acquisition (Cebrian, 2008; Kuhl & Iverson, 1995; Strange, 1995). Listeners would fail to notice the phonetic differences between some L2 sounds when their perceptual systems have been developed to ignore those differences (Escudero, Benders, & Lipski, 2009).

The assumption that L2 perception is influenced by L1 experience also underlies three current prominent paradigms for L2 speech perception: the Speech Learning Model (Flege, 1995, 2003), the Perceptual Assimilation Model's L2 extension (PAM-L2; Best & Tyler, 2007), and the Second Language Linguistic Perception model (L2LP; Escudero, 2005; 2009). The SLM and PAM-L2 posit that initially learners perceive L2 sounds using both their L1 sound categories and L1 perceptual system. The system will cover both L1 sounds and L2 sounds later. The L2LP posits that initially learners fully copy L1 perceptual mappings and L1 perceptual grammar to a new system. L2 sounds will be developed in this new system. The three models indicate that learners either assimilate the L2 sounds to the L1 categories, or not, depending on perceptual similarity between the L2 and L1 sounds. Perceptual similarity is defined differently across the models due to their different postulates about what speech perception is. In the SLM, speech perception is to identify the phonetic cues in the speech stream and map the cues to phonetic categories stored in the mind. Thus perceptual similarity is defined in terms of phonetic distance between the L2 and L1 sounds. The PAM-L2 shares the postulate of the PAM (Best, 1995), which posits that speech perception is to extract the dynamic gestural information (i.e. how the speech signal is formed by articulatory gestures; Browman & Goldstein, 1989) from the speech stream. Perceptual similarity in this model is hence the articulatory difference between the L2 and L1 sounds. The L2LP posits that speech perception is to map the acoustic cues (e.g., spectral and temporal information) to certain discrete and abstract representations in the mind. Thus, perceptual similarity between the L2 and L1 sounds is measured along the acoustic dimension.

Although the SLM, PAM-L2 and L2LP assume that L1 experience shapes the initial state of L2 sound perception, all of them posit that L2 learners can undertake perceptual learning to approximate the native performance in the L2. Based on this postulate, they predict several L2 sounds learning scenarios. It is worth noting that predictions of the SLM are primarily concerned with proficient L2 learners; predictions of the PAM-L2 are mainly

concerned with beginning L2 learners; and the L2LP aims to account for both beginning learners and proficient learners. In addition, predictions of the SLM centre around isolated L2 sounds whereas predictions of PAM-L2 and L2LP are based on L2 sound contrasts. The following section will introduce the predictions of the SLM first, and the predictions of the PAM-L2 and L2LP later.

2.1.2 L2 sounds learning scenarios

The SLM assumes two major learning scenarios for single L2 sounds: the SIMILAR and NEW scenario. The SIMILAR scenario is for L2 sounds that are perceived as similar to L1 sounds. Initially, those sounds are assimilated to the L1 categories and perceived at an allophonic level, in other words, as phonetic realizations of the L1 sounds which do not contribute to meaning distinctions (see also Briere, 1966). Listeners can acoustically differentiate those sounds from the L1 sounds, but they are unable to make use of the differences in speech perception due to "equivalence classification" (see also Morrison, 2002). Sounds in this scenario can never be perceived in a native-like way because no L2 phonetic categories that match the native norms are developed.

In contrast, the NEW scenario is for L2 sounds that are perceived totally different from the L1 sounds. Listeners will not assimilate those sounds to any L1 category. Instead, L2 phonetic categories that match the native norms will be formed. To this end, L2 learners can perceive NEW L2 sounds in a more native-like way in contrast to the SIMILAR L2 sounds. However, the SLM assumes a single phonetic space to cover both L1 and L2 categories. Thus when the newly developed L2 categories are extremely close to existing L1 categories, the listener will strive to maintain a contrast between the categories, like deflecting them, which may result in divergence from the natives.

The PAM-L2 and L2LP predict L2 learning scenarios based on the initial perception of L2 sound contrast. The PAM-L2 describes four possible learning scenarios based on four types of sound contrast resulting from initial L2 perception. The L2LP assumes three scenarios. The two models' predictions will be introduced separately.

The PAM-L2's four scenarios are as follows[1]. SCENARIO ONE: learners assimilate only one sound in the contrast to a specific L1 sound category. The assimilated sound is perceived either as an exemplar or as deviant from the L1 sound at a phonetic level. In either way, the model predicts that the learners will not undertake further perceptual learning because any contrast involving that L2 sound can be discriminated with little difficulty.

[1] The PAM-L2's scenarios are simply called "scenario one, two, three, four" here because the literature does not intend to review the PAM and hence will not use the PAM terms to refer to these scenarios.

SCENARIO TWO: learners assimilate two sounds in the L2 contrast to a single L1 sound category, but one L2 sound is perceived as a better exemplar of the L1 sound than the other. This contrast can be well discriminated as well, but the model predicts that the learners may form an L2 category for the deviant sound at both the phonetic and phonological level. SCENARIO THREE: learners assimilate two sounds in the L2 contrast to the same L1 sound category and both the L2 sounds are perceived as equally good or poor instances of the L1 sound. Learners will find this scenario to be the most difficult one to deal with. The model predicts that a new phonetic category will be formed for at least one L2 sound in the contrast. SCENARIO FOUR: learners do not assimilate any sound in the L2 contrast to any particular L1 category because the L2 sounds are perceived as similar to various L1 sounds. The discrimination of this contrast can vary from poor to very good. The model predicts that one or two new L2 categories may be formed for the contrast, depending not only on the relationship between the L2 sounds and their similar L1 sounds, but also on the overlap of those similar L1 sounds. Table 2.1 summarizes the four scenarios and the predicted level of likelihood for learners to acquire a new L2 category for either phone in the contrast.

The L2LP's three scenarios have similar names to those of the SLM. They are as follows. SIMILAR SCENARIO: L2 sounds in the contrast are assimilated to two different L1 sound categories. Learning sounds in this scenario is assumed to be easy because learners need only replicate their L1 categories and adjust the perceptual mappings from the acoustic cues to the categories. NEW SCENARIO: both L2 sounds in the contrast are assimilated to a single L1 sound. This scenario is assumed to pose the biggest challenge on learners because they have to create both new perceptual mappings and new categories for the L2 contrast. SUBSET SCENARIO: at least one sound in the L2 contrast is acoustically similar to several L1 sounds. Learners in this scenario need to reduce categories copied from the L1 sound system. The learning task is assumed to be easier than that in the NEW scenario and more difficult than that in the SIMILAR scenario. Table 2.2 summarizes the three scenarios and the predicted level of difficulty in acquiring the L2 contrast in a more native-like way.

Table 2.1 The PAM-L2 learning scenarios for L2 sound contrast

Perceptual assimilation pattern	Scenario One	Scenario Two	Scenario Three	Scenario Four
	2L2 onto 2L1; 1L2-0L1 and 1L2-1L1	2L2 onto 1L1, but 1L2 is deviant	2L2 onto 1L1	2L2 onto 0L1
Predicted level of likelihood	Unlikely	Likely	Unlikely	(Uncertain)

Table 2.2 The L2LP learning scenarios for L2 sound contrast

Perceptual assimilation pattern	SIMILAR	NEW	SUBSET
	2L2 onto 2L1	2L2 onto 1L1	1L2-0L1 and 1L2-1L1; 2L2 onto 0L1
Predicted level of difficulty	Less difficult	Most difficult	Medium difficult

2.1.3 Implications

It can be seen that both PAM-L2 and L2LP assume it to be a difficult task to discriminate an L2 contrast with two sounds equally similar to one L1 sound, and to attain this L2 contrast in a native-like way. From the perspective of the PAM-L2, the native-like acquisition is difficult because it is unlikely for learners to form new L2 categories in this scenario. From the perspective of the L2LP, the native-like acquisition is difficult because both new categories and new perceptual mappings between the sound features and the sound categories need to be formed. Both the models also assume that it is relatively easy for the learners to discriminate an L2 contrast with two sounds mapped onto two L1 categories. However, the PAM-L2 assumes that it is difficult to acquire the L2 contrast in a native-like way because no further perceptual learning will be undertaken in this scenario. In contrast, the L2LP assumes the opposite. From its account, only the adjustment of perceptual mapping is needed to acquire the L2 contrast in a native like way in this scenario. This seems to contradict the SLM's prediction that the perception of non-assimilated L2 sounds (NEW scenario) is easier to approximate the native-like level compared to the perception of assimilated L2 sounds (SIMILAR scenario). The contradiction may result from the SLM assuming fewer tasks for learners to approximate native-likeness in the NEW scenario than the L2LP does (Escudero, 2005). The SLM only predicts a category formation task in this scenario. The L2LP predicts that both new categories and new perceptual mappings between the sound features and the sound categories need to be formed. Note that the PAM-L2 does not specify any task regarding the acquisition of L2 sound properties either, which probably causes the slight difference in the prediction made by it and by the L2LP as well.

The three models differ in the degree of explicitness not only regarding how L2 learners attain L2 sound properties, but also regarding the features on which the learners rely to measure the differences or similarities between L2 and L1 sounds. Recall the concept of "perceptual similarity" mentioned in §2.2.1. It is a core concept to the models, as it decides whether the L2 sounds are assimilated or not at the initial stage (Cebrian, 2008). However, only the L2LP specify the method for measuring this similarity: using acoustic features. Neither the SLM nor the PAM-L2 have proposed a consistent and reliable measurement method (Strange, 2007). They predict learners' identification and discrimination performance on L2 sounds using the perceptual assimilation data, but they do not make predictions about the assimilation data itself (Tyler, Best, Faber, & Levitt, 2014; van Leussen & Escudero, 2015). Research on how L2 learners use sound cues may shed further light on the perception models.

Furthermore, although the SLM and the PAM-L2 have not accounted for the effect of using sound cues on L2 perception, they mention that L2 learners may perceive L2 sounds in terms of different properties or dimensions compared to native listeners of the L2, which leads to divergent performance in L2 perception. The L2LP assumes that different cues used by L2 learners are based on the acoustic cues they use in L1 perception. To better understand L2 learners' non-native perceptual performance, the next section will focus on the use of different features by L2 learners in perceiving L2 vowels.

2.2 Perception of Second-language Vowels

2.2.1 Overview

Vowels primarily differ from one another in vowel quality determined by the configuration of the vocal tract during production. The position of the tongue is fundamental to changing the vocal tract shape for the production of vowels. In general terms, the height of the tongue (and that of the jaw) and its position in the front/back dimension vary to create vowels with different qualities. The position (and shape) of the lips is another important parameter in vowel articulation (Cox, 2012; Ladefoged & Johnson, 2014). In some languages and dialects, vowels contrast in length as well, such as Japanese, German, and Australian English (Bohn, 1995; Cox, 2012; Morrison, 2002). Languages and dialects that do not use length to signal phonological vowel distinctions may equip listeners with varied phonetic experience with vowel duration (van Der Feest & Swingley, 2011). For instance, American English vowels are not considered length contrastive but the vowels exhibit different intrinsic duration. Klatt (1976) found that in American English duration could be used as the primary

cue to distinguish stressed and unstressed vowels, the presence or absence of focus and also whether the vowel occurred in a phrase final compared to non-final syllable. Importantly, duration also varies depending on coda context (Hillenbrand, Clark, & Houde, 2000; van Der Feest & Swingley, 2011).

In acoustic signals, vowel quality differences are conveyed as spectral information, which can be identified through examination of frequencies and movement of the first three formants. Length differences are conveyed as temporal (durational) information (Fant, 1971). It is reasonable to hypothesize that listeners from a language with vowel quality contrast and not vowel quantity contrast would use spectral information to discriminate their native vowel sounds. This can be supported by the finding that American English listeners predominantly rely on spectral cues to discriminate their tense-lax vowel pairs (Hillenbrand et al., 2000). However, American English vowel classification experiment using Gaussian models observed an increased identification accuracy when spectral and durational information are combined (Hillenbrand et al., 2000), indicating that duration could also provide some information on vowel identity in American English. This is probably because American English vowel pairs do exhibit a durational difference (as mentioned above). For instance, tense vowels are produced 1.4 times as long as the lax vowels in /hVd/ syllables on average (Hillenbrand, Getty, Clark, & Wheeler, 1995)

It is unclear what cues or cue weightings listeners from a language with both vowel quality and vowel duration contrast would rely on to perceive their native vowels. Classification experiments of AusE vowels using modelling from discrete cosine transform coefficients showed that the duration cue exerts varied influence on AusE vowel identification (Cox, Palethorpe, & Miles, 2015; Watson & Harrington, 1999). Watson and Harrington (1999) found that for the long vowel /iː/, which contrasts with /ɪ/ in both quality and length, removing the duration feature resulted in approximately 7% decrease in classification accuracy, in contrast to 30% decrease in accuracy for the long vowel /ɐː/ which contrasts with /ɐ/ in length only. It is hence reasonable to hypothesize that in Australian English, spectral information is still the primary cue to discriminate vowels contrasting both in quality and length, while durational information plays an important role in discriminating vowels contrasting in length only.

It is also of interest how L2 learners would use the acoustic cues to classify L2 vowel pairs as mentioned in §2.1.3. The PAM-L2 does not subscribe to the idea that acoustic cues form the basic units of speech perception. The SLM only states that L2 learners may use different cue weighting compared to native listeners of the L2. The L2LP specifically

mentions that their strategy may be based on their prior linguistic experience. The following section will review relevant studies.

2.2.2 Cue weighting

A number of L2 vowel perception studies have examined L2 learners' weighting of acoustic cues when they perceive L2 vowel contrasts. The majority of the studies used North American English as the L2 and examined the identification of high front tense-lax vowel contrast /ɪ/-/i/. As mentioned above, native English listeners predominantly rely on spectral information to identify this contrast (e.g., Hillenbrand et al., 2000). Studies are reviewed in two groups according to whether the L2 learners come from a language with phonological vowel duration or not.

The following studies examined L2 learners from an L1 *without* phonological vowel duration. Mandarin learners of English were asked to identify the American English /ɪ/-/i/ contrast in a beat-bit continuum which varied in spectral and durational steps (Bohn, 1995; Flege et al., 1997). The learners overused the duration cue compared to native English listeners. In addition, Mandarin learners who were more experienced with American English made more use of the spectral cues compared to those who were less experienced (Flege et. al., 1997). Spanish learners of English were tested both on American English and Canadian English /ɪ/-/i/ contrast (Bohn, 1995; Kondaurova & Francis, 2008; Morrison, 2009). They predominantly relied on the duration cue as well (but see Escudero, 2005). Spanish learners of Dutch were tested on the Dutch /a:/-/ɑ/ contrast (Escudero et al., 2009). They favoured the duration cue compared to native Dutch listeners for whom the spectral cue was more heavily weighted. Catalan listeners who have varied experience with Canadian English were found to make a greater use of duration to identify English the /ɪ/-/i/ contrast and the duration reliance did not vary with the experience (Cebrian, 2006). Russian learners of English were also examined on the American English /ɪ/-/i/ contrast (Kondaurova & Francis, 2008). Results show that they relied on the duration cue as well.

Studies that examined L2 learners from an L1 *with* contrastive vowel duration have found that Japanese learners of English perceive the Canadian English /ɪ/-/i/ contrast using the same durational information as they use in their own language (Morrison, 2002). German learners' perception of American English /ɛ/-/æ/ contrast were examined (Bohn, 1995). They were found to rely on vowel duration more than the native English listeners would do, but they weighted the duration and spectral cues approximately equally (58.9% reliance on duration; Bohn 1995). Finnish learners of English were both examined and trained on

American English /ɪ/-/i/ contrast (Ylinen et al., 2010). Results before training show that they relied heavily on vowel duration. Results after training show that they were able to use more spectral information than before.

These studies indicate that L2 learners diverge from native listeners of the L2 in relying more on duration cues instead of spectral cues to perceive certain L2 vowel contrast. This is consistent with the prediction of the SLM and PAM-L2. Also, L2 learners rely on duration cues irrespective of their L1 background, i.e. whether the L1 has phonological vowel duration or not. This seems to contradict the prediction of the L2LP, which will be discussed in detail below. Several explanations for the overreliance on durational information among L2 learners are proposed. The following section will also discuss those explanations.

2.2.3 Explanation for durational reliance

L1 transfer and Bohn's (1995) Desensitisation Hypothesis are two common explanations for L2 learners' overreliance on vowel duration[2]. The L1 transfer approach suggests that L2 learners directly transfer their experience with vowel duration in L1 to L2 vowel identification, i.e., they use vowel duration as a cue in their L1 and hence also use the cue in L2 (e.g., Kondaurova & Francis, 2008). In contrast, Bohn's (1995) Desensitisation Hypothesis suggests that durational reliance is a general perceptual strategy that all L2 learners use independently of L1.

A challenge for the L1 transfer approach is to explain the cases of L2 learners who do not use durational information to discriminate vowels in their L1, e.g., Spanish, Russian and Mandarin learners but can nonetheless use duration as a cue in the L2. Kondaurova & Francis (2008) argue that Spanish and Russian learners have allophonic experience with vowel duration in their L1 which can be transferred to L2 vowel identification: both of them use vowel duration as cue to lexical stress. Flege et al. (1997) indicate that Mandarin learners have a reliable experience using vowel duration to differentiate tones. However, questions remain about how learners can directly transfer experience in one domain, i.e. coda voicing, lexical stress, tones, to another domain, i.e. vowel identification (Escudero et. al, 2009).

Bohn (1995)'s Desensitisation Hypothesis does not assume the use of duration cue among L2 learners to be a result of prior linguistic experience. Instead, the duration cue is considered to have universal saliency so that it can be accessed without prior experience. In

[2] There are also other explanations for L2 learners' reliance on vowel duration, such as the possible effect of EFL instruction where vowels are often described as short or long without reference to quality differences (e.g., Cebrian, 2006). Those explanations have not been accounted for in detail in this thesis because the current study aimed to explore the L1 transfer, the Desensitisation Hypothesis, and the L2LP explanation.

contrast, spectral cues are more difficult to access in the absence of experience. Thus listeners would rely on the durational information to discriminate L2 vowel contrasts when they are not able to detect spectral differences because L1 acquisition has desensitized them to those differences. This hypothesis seems to explain the durational reliance among L2 learners who have no phonological experience with vowel duration in L1. However, Escudero & Boersma (2004) express doubt about the hypothesis's core idea that vowel duration is universally salient. They argue that vowel duration is not special. L2 learners who have no phonological experience with vowel duration tend to rely on this cue just because they "have a 'blank slate' on this dimension" and hence can easily "form categories along the dimension" (Escudero et al., 2009, p. 463). The explanation proposed by Escudero & Boersma (2004) was incorporated into the L2LP model (Escudero, 2005). The remainder of this section will present their explanation under the L2LP framework.

As mentioned in §2.1.1, the L2LP assumes a full copy of L1 perceptual system and a full copy of L1 perceptual learning mechanism at the initial state of L2 learning. Therefore, L2 learners perceive L2 vowels along the acoustic dimension they use to perceive L1 vowels and are able to create new mappings between acoustic cues and categories. When L2 learners perceive an L2 vowel pair contrasting mainly in the spectral dimension but also exhibiting some durational differences, they map the spectral and durational information of the vowel pair onto their own spectral and durational dimension separately.

For learners who have no phonological experience with vowel duration, they have no categories on the durational dimension while their spectral categories have been well established in L1 acquisition. It is easier to form two new categories for the long-short durational contrast on a blank slate rather than adjust existing categories or generate new categories on a well categorized spectral dimension. Thus learners pay special attention to the durational information and rely on the "long" and "short" categories they create to identify the vowel pair at the initial stage.

For learners who already have two phonological durational categories, there is also a high probability that they rely more on the durational categories compared to spectral categories at the initial stage. There may be three possible spectral mapping scenarios for the L2 vowel pair: two vowels spectrally map onto one L1 vowel (NEW scenario), or at least one of the two vowels maps onto several L1 vowels (SUBSET scenario), or the two vowels map onto two L2 vowels separately (SIMILAR scenario). Only in the last scenario, can the listener differentiate the L2 vowel pair using the spectral cue. In contrast, there is only one mapping scenario on the durational dimension: two vowels onto "long" and "short"

categories respectively. Thus it is more reliable for the L2 learners to use the duration cue to discriminate the L2 vowel pair at the initial stage.

The L2LP's explanation can account for the cue weighting change among L2 learners as well. Some studies above show that more experienced L2 learners would make greater use of spectral information to identify the L2 vowel pair. The L2LP suggests that it is due to the mappings adjustment and/or categories creation along with L2 perceptual learning.

2.2.4 Implications

A number of studies have examined how L2 learners weight acoustic cues when they are perceiving L2 vowel pairs that contrast mainly in spectral information. Of particular interest is that some L2 learners show overreliance on vowel duration even though they do not use the cue to signal vowel contrast in their L1. Three explanations are introduced above: L1 transfer, Bohn's Desensitisation Hypothesis, and the L2LP model. All exhibit two general perspectives: cross-linguistic (L1 transfer) and developmental (Bohn's hypothesis and L2LP). The cross-linguistic perspective attributes L2 learners' durational reliance to their allophonic experience with vowel duration in L1. The developmental perspective indicates that the L2 learners form phonological duration categories (i.e. long or short) along with their exposure to the L2, due to the duration cue's saliency (Bohn's hypothesis) or due to they have a blank slate on the duration dimension (L2LP).

The two perspectives differ in the degree to which the L2 learners can use vowel duration in L2 perception. The developmental perspective suggests that the L2 learners can and would use vowel duration in a more native-like way than the cross-linguistic perspective suggests, because forming categories for a length contrast is a typical L1-like acquisition strategy (Escudero, 2005). Thus, the two perspectives can be examined through comparing the L2 learners and native listeners' perception of vowel duration.

However, as mentioned above, the majority of prior identification studies used North American English as L2 and examined the tense-lax vowel pair which contrast mainly in vowel quality. Native English listeners would primarily use spectral cue in those tasks, so that their perception of vowel duration can barely be observed. Furthermore, the examined vowel pairs in previous studies were usually varied in equal spectral and durational steps, which is a suitable design for a cue weighting study, but not satisfactory for observing L2 learners and native listeners' interpretation of vowel duration differences without corresponding vowel quality differences. Thus there are almost no experimental data available that can be used to compare L2 and native English listeners' perception of vowel

duration while controlling for both duration and spectral information[3]. This study hopes to fill the gap.

To enable a full observation of how L2 learners and native listeners use vowel duration in vowel identification, Australian English, in which the weighting between spectral and durational cues varies across different vowel pairs (Cox et al., 2015), is chosen as the L2. Mandarin listeners who do not use vowel duration to distinguish L1 vowels are chosen as the L2 learners. The following section will discuss the AusE vowel pairs selected for this study (§2.3.1), the context in which they will occur (§2.3.2) and how Mandarin listeners may respond to these vowels according to the L2LP (§2.3.3).

2.3 Mandarin Listeners and Australian English Vowels

2.3.1 Australian English vowels

Australian English has 18 stressed vowels, 12 of which are the monophthongs: /ɪ e æ ɐ ɔ ʊ i: e: ɐ: o: ʉ: ɜ:/. Most monophthongs are distinguished in both vowel quality and vowel length (Cox, 2012). There are six short vowels and six long vowels, and short vowels are approximately 60% the length of the long vowels in the hVd context (Cox, 2006). For the present study we chose two vowel pairs to examine L2 and native AusE listeners' perception of vowel duration: /ɐ/-/ɐ:/ (*hut-heart*) and /ɔ/-/o:/ (*hot - hort*).

/ɐ/ and /ɐ:/ are duration contrastive in non-rhotic AusE (Cox, 2015; Harrington, Cox, & Evans, 1997). X-ray data (Bernard, 1970a) and acoustic studies (Bernard, 1970b; Cox, 2006; Watson & Harrington, 1999) have confirmed the absence of spectral differentiation for this vowel contrast. A perceptual task using this vowel pair requires categorization of the vowels based on duration only. In contrast, /ɔ/ and /o:/ have both spectral and durational differences. They have the same degree of backness, but /ɔ/ is phonetically lower than /o:/ (Cox, 2006). A perceptual task using this vowel pair requires the listeners to integrate different acoustic cues. The /ɐ/-/ɐ:/ contrast is intended to investigate if Mandarin learners of English can perceive

[3] There is a study that examined the acquisition of Swedish vowel duration by speakers from different L1 background: native Swedish speakers, and English, Estonian, Spanish L2 learners of Swedish (McAllister, Flege, & Piske, 2002). The study tested those speakers' perceptual sensitivity to Swedish words containing four pairs of quantity contrasted vowels: two of the vowel pairs (/ø:/-/ø/, /ɛ:/-/ɛ/) have only duration distinction and the other two (/ʉ:/-/ʉ/, /a:/-/a/) have both vowel quality and duration distinction. Half of the stimuli were created by producing real Swedish words containing the eight vowels. The other half of the stimuli were created by producing the real words containing the long vowels with the corresponding short vowels and vice versa. Participants were asked to determine whether the word they heard was phonologically correct (words with right vowel length) or not (words with the opposite vowel length). It was found that the native Swedish listeners and Estonian L2 learners outperformed English L2 learners, and English learners outperformed Spanish L2 learners.

AusE vowel duration in a native-like way. The /ɔ/-/oː/ contrast is intended to compare the effect of duration on Mandarin listeners' and native AusE listeners' perception.

All vowels will be included in syllables and presented as words. This is because the current study focuses on listeners' perception of vowel duration and vowel duration is sensitive to the context (Cox, 2006; Harrington & Cassidy, 1999). The following section will discuss how coda voicing influences AusE vowel duration.

2.3.2 Voicing effect

Although AusE vowels have inherent duration, the differences between long and short vowels become relative when the vowel is contextualized (Cox, 2006; Harrington & Cassidy, 1999). Coda voicing is one of the contexts that exerts a strong influence on vowel duration. Generally, vowels preceding voiced consonants are longer than when they precede voiceless consonants. The length ratio is approximate 1.8:1 for English vowels in voiced- and voiceless-coda syllables (Raphael, 1972) but this ratio is dependent on whether the vowel is an inherently long or short vowel. Cox and Palethorpe (2011) and Cox et al. (2015) showed that long vowels lengthen more than short vowels in voiced coda contexts. The vowel duration is thus used as a cue to coda voicing but the ratio of vowel duration in the voiced to voiceless contexts varies according to whether the vowel is long or short.

It is of interest then whether coda voicing influences listeners' duration categorization. The hypothesis would be: vowels followed by voiced coda need to be longer to be categorized as long vowels than followed by voiceless coda and vice versa. Comparing the voicing effect on Mandarin listeners and native AusE listeners can also show whether they perceive vowel duration in a similar way.

2.3.3 Mandarin listeners' perception of AusE vowels

The Mandarin Chinese vowel system comprises a range from three to 12 monophthongs according to different analyses (Flege et al., 1997). Five or six vowels are most commonly recognized (Li & Thompson, 1981; Maddieson & Disner, 1984). This study follows Howie's (1976) analysis using a six vowel system /i y a ɤ u ə/. Duration is not used to distinguish Mandarin vowel pairs (Bohn, 1995; Flege et al., 1997), but it may act as a cue to differentiate the second and third tone (Blicher, 1988). Mandarin vowels occur either in open syllables or preceding the consonant /n/, so the Mandarin listeners have no experience with coda voicing contrast (Flege et al., 1997).

Few perception studies have examined native Mandarin listeners or Mandarin L2 listeners' perception of AusE vowels. Thus the current study compares Mandarin and AusE vowels in the vowel map and presents predictions about how Mandarin learners of English

will perceive AusE vowel pairs /ɐ/-/ɐː/ and /ɔ/-/oː/ according to the L2LP. Figure 2.1 shows the vowel chart of Standard AusE, which is regarded as "the dominant variety of English spoken in Australia" (Cox, 2006). Figure 2.2 shows the vowel chart of Mandarin Chinese as they are pronounced in Beijing (Lee & Zee, 2003).

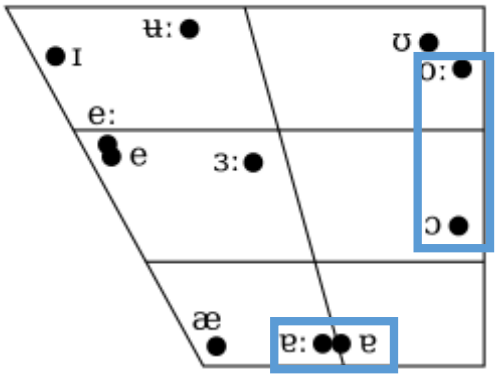


Figure 2.1 Monophthongs of AusE (Cox, 2012, p. 159)

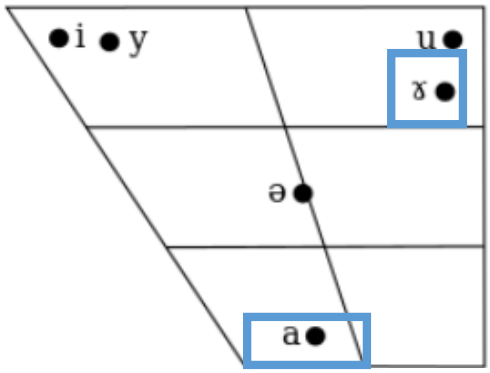


Figure 2.2 Monophthongs of Mandarin Chinese (Lee & Zee, 2003, p.110)

It can be seen from the chart that the AusE vowel pair /ɐ/ and /ɐː/ are spectrally similar to Mandarin /a/ when viewed from the perspective of general phonetic positioning in the vowel space. In addition, there are no other Mandarin vowels similar to /a/. Thus Mandarin learners of English may assimilate AusE /ɐ/ and /ɐː/ into one L1 category /a/. This is the NEW scenario in L2LP. Mandarin learners will not discriminate this AusE vowel contrast by spectral information at the initial stage and they are not able to discriminate the two L2 vowels from their L1/a/. As the L2 vowel pair contrasts in duration, Mandarin learners can establish native English-like duration categories to identify the two vowels. When the learners are more experienced with AusE, they may establish a spectral category for the two L2 vowels to separate them from the L1 /a/, but they will still primarily rely on duration to discriminate this L2 contrast.

It is more complicated to predict how AusE /ɔ/-/oː/ will be assimilated to Mandarin vowel categories. One possibility is that both /ɔ/-/oː/ are assimilated into Mandarin /ɤ/. This is also the NEW scenario in which learners have difficulties detecting spectral differences between /ɔ/-/oː/. In this scenario, Mandarin listeners will rely on duration to identify the two vowels at the initial stage. They may generate another category or split existing category for one of the two L2 vowels when their experience with AusE increases. Another possibility is that /oː/ is mapped onto Mandarin /ɤ/ and/or Mandarin /u/, and /ɔ/ is mapped onto no L1 category. This is the SUBSET scenario, in which listeners can discriminate the two vowels spectrally from the initial stage, but they would still be influenced by the duration cue (see§2.2.3).

2.4 Current study

The current study will examine Mandarin L2 learners and native AusE listeners' perception of AusE vowel pairs /ɐ/-/ɐː/ and /ɔ/-/oː/ in both voiced and voiceless contexts. The main aim is to see whether Mandarin learners perceive vowel duration in categories and whether their duration-based categories resemble the native AusE listeners'. Specific research questions for each vowel pair are as follows:

Research Question 1: Do Mandarin and AusE listeners use a duration-based category to perceive the AusE vowel contrast /ɐ/ - /ɐː/? If the answer is positive, are the category boundaries similar or different between Mandarin and AusE listeners in terms of location and steepness?

Research Question 2: Do Mandarin and AusE listeners use a duration-based category boundary to perceive the AusE vowel contrast /ɔ/ - /oː/? If the answer is positive, are the category boundaries similar or different between Mandarin and AusE listeners in terms of location and steepness?

Research Question 3: If Mandarin and AusE listeners show duration-based categories, will coda voicing influence the location of their category boundaries?

Chapter 3 Method

3.1 Participants

Sixty students from Macquarie University were recruited as participants: 30 Mandarin learners of English and 30 native Australian English listeners. Given the time constraint for the project and to make the two language groups comparable, the current study recruited female participants only. The recruitment criteria for Mandarin listeners were: native Mandarin listeners who arrived at Australia after the age of 12 and have studied and lived in Australia for at least six months. The residence length cut-off was set at six months because the study needs the L2 learners to have undergone some L2 perceptual learning to establish the duration categories if they can. Six months seems to be the most beneficial period for the L2 learners' perceptual learning, because 1) compared to late L2 learners with 0-6 months of experience living in an L2 environment, those with 6 -12 months of experience were reported to undergo significant L2 perceptual learning; 2) after 6-12 months' immersion, little perceptual benefits were observed from additional experience for most late L2 learners (Best & Tyler, 2007; Flege & Liu, 2001). The second finding may not apply to L2 learners who extensively interact with native speakers of the L2 (Flege & Liu, 2001), but it does not influence the conclusion that the first six months immersion in the L2 environment is critical for most L2 learners. The recruitment criteria for native AusE listeners were: native AusE speakers born to AusE-speaking parents. This is because AusE speakers who were born and raised in Australia were found to be less sensitive to changes in vowel duration if they had at least one parent speaking another dialect of English which does not have vowel length contrasts (Chen, Xu Rattanasone, Cox, & Demuth, 2015).

The Mandarin group comprised participants aged from 19-30 years old (*mean=23.9*, *sd=3.0*). All the participants in this group were born to Mandarin-speaking parents in Mainland China: 12 were born in northern cities (i.e. Beijing, Shenyang etc.) and 18 were born in southern cities (i.e. Nanjing, Guangzhou etc.). They started to learn English at an average age of 9.1 years old (*sd=2.7*) in a classroom setting in China, and all came to Australia after the age of 16 (*mean=21.4, sd=3.3*). Before coming to Australia, only two participants had lived in another foreign country for more than half a year: one stayed in Singapore for one year at the age of 18; the other lived in America for 0.7 year at the age of 22. At the time of testing, the 30 Mandarin participants had studied in Sydney for a range of 0.5 year to 7 years (*mean=2.4, sd=1.7*). Twenty-one of them (70%) self-reported to use English for less than 50% of daily communication. The remaining nine (30%) reported a

frequency between 50% to always. In contrast, twenty-three of them (77%) estimated their listening comprehension proficiency in Australian English as intermediate or above (on a seven-level proficiency scale). The remaining seven (23%) self-reported to be below the intermediate level. Table 3.1.1 presents a summary of the characteristics of the 30 Mandarin participants.

The native AusE group aged from 18-52 years old (*mean=22.6, sd =8.1*). All the participants were born and raised in Australia by parents who were also native Australian English speakers. Only two participants had the experience living in other English-speaking countries: one lived in America for 0.5 year at the age of 23 and in Britain for one year at 24; the other lived in Canada for 0.5 year at the age of 20. At the time of testing, all of them were undertaking undergraduate study in the Department of Linguistics. Twenty-three participants (76%) had learned another language/languages for more than six months, including French, Italian, Japanese etc., but no one had learned Mandarin as an L2. All of them reported an infrequent use of the second language in daily life (i.e., no use or less than 25% of daily communication). Only two participants had the experience living in an L2 country: one lived in France for 0.8 year at the age of 22; the other lived in Japan for one year at the age of 16. Table 3.1.2 presents a summary of the characteristics of the 30 Australian English participants.

All 60 participants self-reported normal hearing and having no speech disorders. After the experiment, 23 participants received research credit to meet their unit requirement (i.e., units of Linguistics). The remaining were paid $20 each as compensation for their time[4].

[4] Approved by Macquarie University Human Research Ethics Committee (ref: 5201600131).

Table 3.1.1 Characteristics of the 30 female Mandarin participants

Characteristics*	**Min**	**Max**	**Mean**	**Standard Deviation**
Chronological age (years)	19.0	30.0	23.9	3.0
Age of starting to learn English (years)	3.0	14.0	9.1	2.7
Age of arrival in Australia (years)	16.0	29.0	21.4	3.3
Length of residence in Australia (years)	0.5	8.0	2.6	2.0
Length of study in Australia (years)	0.5	7.0	2.4	1.7

% self-reported use of English (number and percentage of participants)	**Never**	**<25%**	**25%-50%**	**50%-75%**	**>75%**	**Always**
	0	8 (27%)	13 (43%)	7 (23%)	2 (7%)	0

Self-reported listening comprehension proficiency of Australian English (number and percentage of participants)	**Not at all**	**2**	**3**	**Medium**	**5**	**6**	**Very Good**
	1 (3%)	1 (3%)	5 (17%)	10 (33%)	8 (27%)	5 (17%)	0

Place of birth in China** (number and percentage of participants)	**Northern cities**	**Southern cities**
	12 (40%)	18 (60%)

Live in another foreign country for more than 0.5 year	**Number of participant**	**Country**	**Age/length of living**
	1	Singapore	18 years old / 1 year
	1	America	22 years old / 0.7 year

*Note: * See appendix A for a full list of the birth cities and the distribution of the participants.*

Table 3.1.2 Characteristics of the 30 female Australian English participants

Chronological age (years)	**Min**	**Max**	**Mean**	**Standard Deviation**
	18.0	52.0	22.6	8.1

Learn a second language (L2) for more than 0.5 year (number and percentage of participants)	**No L2**	**One L2**	**Two L2s**	**Four L2s**
	7 (24%)	21 (70%)	1 (3%)	1 (3%)

% self-reported use of L2 (number and percentage of participants)	**Never**	**<25%**	**25%-50%**	**50%-75%**	**>75%**	**Always**
	20 (67%)	10 (33%)	0	0	0	0

Live in another foreign country for more than 0.5 year	**Number of participant**	**Country 1**	**Age/length**	**Country 2**	**Age/length**
	1	America	23 years old / 0.5 year	Britain	23 years old / 1 year
	1	France	22 years old/ 0.8 year		
	1	Japan	16 years old/ 1 year		
	1	Canada	20 years old/ 0.5 year		

Place of birth in Australia (number of participants)	**NSW**		**WA**		**NT**	**VIC**	**QL**
	Sydney	Other	Perth	Port Hedland	Darwin	Geelong	Mount Isa
	23	2	1	1	1	1	1

3.2 Materials

The experiment was a two-forced-choice identification task. It examined how variations in vowel duration in the AusE vowel pairs /ɐ/-/ɐː/ and /ɔ/-/oː/ affects listeners' categorization, both in /hVd/ and /hVt/ contexts. The stimuli for the task were duration-manipulation versions of natural recordings, rather than synthesized from scratch. A female native speaker of AusE produced each of the vowels /ɐ/, /ɐː/, /ɔ/, and /oː/ in both /hVd/ and /hVt/ context. The vowel portion of each token was manipulated in 11 duration steps, yielding 11 stimuli containing duration-edited vowels in /hVd/ or /hVt/ syllable form. On each trial of the task, participants were presented with two target words exemplifying a vowel contrast on the screen, and with a stimulus generated from the token containing either vowel in that contrast over the headphone. Their task was to classify the vowel in the stimulus as either vowel in that vowel pair, and responded by referring to the target word. All the participants were familiarised with the target words for the vowels prior to the identification task, making sure that their choice of the lexical item properly reflects their choice of the vowel. The remaining of this section will give more details about the natural tokens and target words for the vowels (§3.2.1), the stimuli for the actual identification task (§3.2.2), and the materials for the familiarisation purpose (§3.2.3).

3.2.1 Natural tokens and target words

Eight naturally produced tokens containing the vowels /ɐ/, /ɐː/, /ɔ/, and /oː/ in /hVd/ and /hVt/ form were selected from the citation form productions of a 20-year-old Australian female university student (Cox & Palethorpe, 2010). The student was born and raised in Sydney's North West. Her parents were also native Australian English speakers and both undertook professional occupations. The recording took place in the recording studio of the Department of Linguistics at Macquarie University. The speaker was asked to read three lists of words containing the 18 stressed vowels of AusE in a range of consonantal contexts (see Cox & Palethorpe, 2011). Each list was produced 3 times using separate randomisations and were recorded using Cool Edit on a Pentium 4 PC (M-Audio delta 66 sound card) at 44.1kHz sampling rate via an AKG C535 EB microphone. The eight tokens were selected on the basis of having similar recording quality and the least degree of glottalisation, so that listeners cannot rely on the recording or glottalisation information to identify the stimuli created from these tokens. The spectrograms of the eight natural tokens are presented in Appendix B.

The elicitation words for the eight natural tokens are presented in Table 3.2.1.1. They were: "hud" for /hɐd/, "hard" for /hɐːd/, "hod" for /hɔd/, "horde" for /hoːd/, "hut" for (/hɐt/),

“heart” for /hɐːt/, “hot” for /hɔt/, and “hort” for /hoːt/. This set was also used as the target words representing target vowels in the identification task. It is noteworthy that there were three non-words in the lexical set: hod, hud, and hort. The three non-words have plausible spelling so that they can represent the intended vowels as well as the other five real words can. However, lexical effects like the Ganong effect (Ganong, 1980) might be observed when the identification task was performed. Participants might tend to identify the stimuli as real words rather than non-words. This perceptual bias towards lexical items will be accounted for in the discussion section.

Table 3.2.1.1 Target words for the vowels

	Length (/ɐ/ - /ɐː/)	**Length-Spectral (/ɔ/ - /oː/)**
voiced (/hvd/)	hud - hard	hod - horde
voiceless (/hvt/)	hut - heart	hot - hort

3.2.2 Stimuli for identification task

The eight natural tokens were manipulated using Praat (Version 5.4.17; Boersma & Weenink, 2015) to create stimuli for the identification task. There were four steps to the manipulation process: 1) vowel duration manipulation, 2) extraction and onset/coda normalization, 3) F0 normalization, and 4) intensity normalization. The manipulation resulted in 11 stimuli with vowel duration ranging from 85ms to 335ms for each token (i.e. “hud”, “hard”, “hod”, “horde”, “hut”, “heart”, “hot”, “hort”). The four manipulation steps are clarified as follows:

Step 1: vowel duration manipulation

First, the vowel portion of each token was manually annotated for later manipulation. The spectrogram was set to view a range of 0-4000Hz and window length was set to 0.005s. Automatic formant tracking was achieved by the LPC formant tracking algorithm with a standard setting (i.e. maximum formant: 4000Hz, number of formants: 4.0, window length: 0.05s, dynamic range: 30.0dB, dot size: 1.0mm). The beginning of the vowel was labelled at the onset of periodicity. The end of the vowel was labelled at where the energy of the second and third formant ceased. The labelled vowel duration is presented in Table 3.2.2.1.

Table 3.2.2.1 Vowel duration of the eight natural tokens

Token	**Vowel**	**Vowel duration**	**Token**	**Vowel**	**Vowel duration**
hud	/ɐ/	152ms	hut	/ɐ/	103ms
hard	/ɐː/	335ms	heart	/ɐː/	229ms
hod	/ɔ/	138ms	hot	/ɔ/	124ms
horde	/oː/	327ms	hort	/oː/	245ms

Second, duration of each labelled vowel was varied in 11 steps with the endpoints 85ms and 335ms and an equal interval of 25ms between the steps, yielding 88 sound files (11variation × 8 tokens = 88 sound files). The interval between steps was set to 25ms because 25ms may be considered the just-noticeable difference (JND) for English listeners (Klatt, 1976; Kondaurova & Francis, 2008). The endpoints, 85ms and 335ms, were determined based on the original vowel duration, which ranged from 103ms to 335ms (Table 3.2.2.1). 85ms instead of 103ms was chosen as the first step to allow for a whole number of steps. The duration manipulation was realized through the manipulation function in Praat, which can lengthen or shorten a labelled object to a target duration by adding relative duration points at the object's boundaries (relative duration points = target duration/original duration) and resynthesized sound using the overlap-add method. For example, the vowel /ɐ/ in the token "hud" has a duration of 152ms. Praat lengthened it to 335ms by adding relative duration points "2.2" (335ms/152ms = 2.2) at its boundaries.

It is worth mentioning that the current manipulation method varied the vowel's duration but maintained the vowel's target-to-transition (i.e., onglide, offglide) ratio. This method was chosen over another widely adopted method, which manipulates vowel duration by splicing in or cutting off some target components and hence will change the total vowel duration as well as its target-to-onglide/offglide ratio. There are two theoretical reasons for the current study to maintain the ratio. Firstly, short vowels are contrasted with long vowels in production, not only in that their total duration is shorter, but also in that they have smaller target-to-offglide ratio than the long vowels do (Felicity Cox, 2006; Lehiste & Peterson, 1961). Thus, both the vowel's total duration and the target-to-transition ratio can be important perceptual cues to the vowel's length contrast. Changing the duration and ratio at the same time will pose difficulties on the observation of single effects from either of the two factors (i.e. vowel duration and target-to-transition ratio). Furthermore, the target-to-offglide ratio is argued to influence listeners' perception of long and short vowels by providing information on the timing of articulation gestures (Pycha & Dahan, 2016). Thus any results attributable to

a change of the target-to-offglide ratio are due to the listeners' reliance on both durational (i.e. timing) and spectral cues (i.e. articulation gestures). As the present study aimed to compare the durational effect with the spectral effect, it was important to control the ratio factor which shows a combination of those effects.

Additionally, from a practical perspective, it was impossible for the current study to adopt the splicing method which varies the vowel duration by manipulating the target components alone. Some vowels' transitional duration (i.e. the combined onglide and offglide duration) exceeded a particular target duration (which is for a whole vowel) making it impossible to reduce the vowel duration sufficiently because the combined onglide and offglide duration exceeded the minimum vowel length of 85 ms. For example, /ɐː/ in the token "hard" has an original duration of 335ms. The target components of /ɐː/ in /hVd/ context generally take up 62% of the whole vowel (Cox, 2006), so the target duration is approximately 335ms × 62% = 208ms and the transitional duration is 335ms – 208ms = 127ms. This means when all the target components were excluded, the vowel still has a duration of 127ms. Then it is impossible to shorten /ɐː/ in /hVd/ to 85ms or 110ms (i.e. the first two steps of the 11 duration steps) by manipulating its target components only.

Step 2: extraction and onset/coda normalization

In the sound files generated from *Step 1*, silence of varied length existed before and after the sounds (i.e. /hVd/ or /hVt/ syllables), as the eight source tokens were recorded with some silence. The sound portion of each sound file was extracted to be the final stimuli. However, the extraction did not start from the right beginning (i.e. where the waveform and spectrogram started) to the right ending of each sound (i.e. where the waveform and spectrogram ended) for the purpose of generating stimuli with normalized onset and coda duration. In each sound file, the starting extraction boundary was set at 77ms before the vowel began (see *Step 1* for criteria). The ending boundary was set at different position after the vowel ended for different set of sound files (as presented in Table 3.2.2.2): 137ms for sound files based on the "hard" and "hud" token, 253ms for sound files based on the "heart" and "hut" token, 165ms for sound files based on the "horde" and "hod" token, and 274ms for sound files based on "hort" and "hot" token. As a result, the 88 extracted sounds had onset (i.e. /h/) of the same length. Sounds containing a contrasted vowel pair in the same context had a comparable coda (i.e. /d/ or /t/) length.

Table 3.2.2.2 stimuli's onset and coda duration

Voicing	Vowel	Source Token	Onset /h/	Coda /d/ or /t/
Voiced /hVd/	/ɐː/	hard	77ms	139ms
	/ɐ/	hud	77ms	139ms
	/oː/	horde	77ms	165ms
	/ɔ/	hod	77ms	165ms
Voiceless /hVt/	/ɐː/	heart	77ms	253ms
	/ɐ/	hut	77ms	253ms
	/oː/	hort	77ms	274ms
	/ɔ/	hot	77ms	274ms

Stimuli's onset duration was normalized to allow for a consistent reaction time measurement (see §3.3 for details). The duration was set to 77ms because all the stimuli needed to start from the sound directly (instead of silence) and the shortest onset duration for those sounds was 77ms. The coda duration of the stimuli with contrasting vowels was normalized because these stimuli were played in the same block during the identification task (see §3.3 for details). The listeners were expected to rely on the stimuli's vowel differences when making their selection rather than any other cues, e.g., coda difference. The specific value was decided based on the shortest coda of the set.

Step 3: F0 normalization

The 88 stimuli generated from *Step 2* were then normalized in F0, as F0 can influence the perception of duration. Vowels are perceived longer on a dynamic F0 contour (e.g., falling or rising) compared to on a static or flat F0 contour (e.g., Yu, 2010). Vowels with a high mean F0 are also perceived longer than with a low F0 (e.g., Gandour, 1977; Yu, 2010). Thus both the F0 pattern (i.e. contour) and F0 height (i.e. mean) needed to be equalized for vowels in the 88 stimuli.

An F0 contour falling from 250Hz to 190Hz was interpolated in the vowel portion of each stimulus after the vowel's original F0 contour was removed[5]. 250Hz-190Hz falls into the general fundamental frequency range for adult females (i.e. 165Hz-255Hz, Traunmüller & Eriksson, 1995). The contour's mean frequency (250Hz + 190Hz)/2 = 220Hz also equals to the mean fundamental frequency for females aged 20+ (Stoicheff, 1981; Traunmüller & Eriksson, 1995).

[5] A dynamic contour instead of a static contour was adopted because the latter made the stimuli sound unnatural according to the pilot study.

Step 4: intensity normalization

Finally, the average intensity for each stimulus was set to 65dB. The onset /h/ of each stimulus was faded in and the coda /d/ and /t/ were faded out. That is, the first half of /h/ was multiplied with a (1-cos (x))/2 function and the first half of /d/ or /t/ was multiplied with a (1+ cos (x))/2 function (Boersma & Weenink, 2015).

3.2.3 Materials for familiarisation purpose

Prior to the actual identification task, participants were familiarised with target words for vowels to make sure they had established a proper connection between the lexical form and the vowel categories. They were presented with the eight words (i.e. hard, hud, horde, hod, heart, hut, hort, and hot) both orthographically and audibly (see §3.3 for details). It is worth mentioning that the audio tokens used here were not the eight natural tokens the 88 stimuli were based on. Tokens for the familiarisation purpose were produced by another female native AusE speaker, who was also a university student, born and raised in Sydney's North West by AusE speaking parents, but was 19 years old at the time of recording (the first speaker was 20). The recording environment and equipment were all the same for the two speakers.

The reason for using different sources to generate sound files for the familiarisation and actual test purpose was to decrease the possibility that listeners would base their categorization during the test on cues in the natural recordings that they were exposed to in the familiarization phase. More specifically, if the same set of eight natural tokens were used at both stages (i.e. familiarisation and identification), listeners may detect some uncontrolled differences between the natural tokens and relied on those differences to identify the stimuli generated from them. The task would then be one of "trace back the source of the stimuli", rather than "identify the vowel category".

Tokens for the familiarisation purpose were normalized in terms of F0 (250-190Hz) and intensity (65dB). No further modifications have been made because the tokens need to be exemplars of the target vowels. The following section will describe the specific procedures of data collection.

3.3 Data collection

Participants were tested individually in the Speech Perception Lab at Macquarie University. Both AusE natives and Mandarin learners of English were tested by a native Mandarin speaker who gave the instructions in English. The participants were told that they would listen to AusE sounds and perform a two-forced-choice identification task.

The experiment was implemented on a Mac using PsyScope X B77 (Cohen, MacWhinney, Flatt, & Provost, 1993). Audio stimuli were played via Sennheiser HD380 Pro headphones. Response choices (i.e. target words) were displayed on the screen: one at the left side and the other at the right side, corresponding to the "a" and "l" keys on the keyboard. The 88 stimuli were tested in four blocks, examining the two vowel contrasts /ɐ/- /ɐː/ and /ɔ/- /oː/ in two contexts respectively. In each block, the response choices remained unchanged. Table 3.3.1 shows how the stimuli were grouped and the response choices for each block.

Table 3.3.1 Stimuli presentation and response choices

Blocks	Context	Vowel contrast	Response choices	Stimuli
1	/hVd/	/ɐ/- /ɐː/	hud - hard	11 from the token "hud" 11 from the token "hard"
2		/ɔ/-/oː/	hod - horde	11 from the token "hod" 11 from the token "horde"
3	/hVt/	/ɐ/- /ɐː/	hud - heart	11 from the token "hot" 11 from the token "hort"
4		/ɔ/-/oː/	hot - hort	11 from the token "hot" 11 from the token "hort"

Listeners passed through three stages in each block: familiarisation, practice test and actual test. Take the block examining /ɐ/- /ɐː/ contrast in /hVd/ context as an example. ***At the stage of familiarisation***, listeners first saw a cross at the centre of the screen. The cross disappeared 500ms later, followed by two response labels "hard" or "hud" displayed at either right or left side on the screen and its corresponding natural token (as mentioned in §3.2.3) presented over the headphone at a 1000ms delay. The natural token was played only once. When the sound finished, listeners pressed any key to move to the other word. Each word was presented three times in sequence. ***At the stage of practising,*** after the cross disappeared, "hard" and "hud" were displayed on the screen simultaneously and the natural token of either word was played once at a 500ms delay. The listeners needed to choose the word they heard by pressing the "a" or "l" key on the keyboard representing the left or right of the screen respectively. The program then gave feedback by removing the incorrect word from the screen and playing the correct word for one more time. There were six trials at this stage, with each word played three times in a random order. The actual test was quite similar to the practice test, but used the edited stimuli instead of natural tokens and had more trials without giving feedback. ***At the stage of actual testing,*** listeners were presented with "hard" and "hud" orthographically on the screen, and with one of the 22 stimulus items over the headphones. They were asked to respond as soon as possible. The 22 stimuli were played four times in a random order and hence there were 88 trials in each block. The whole block lasted

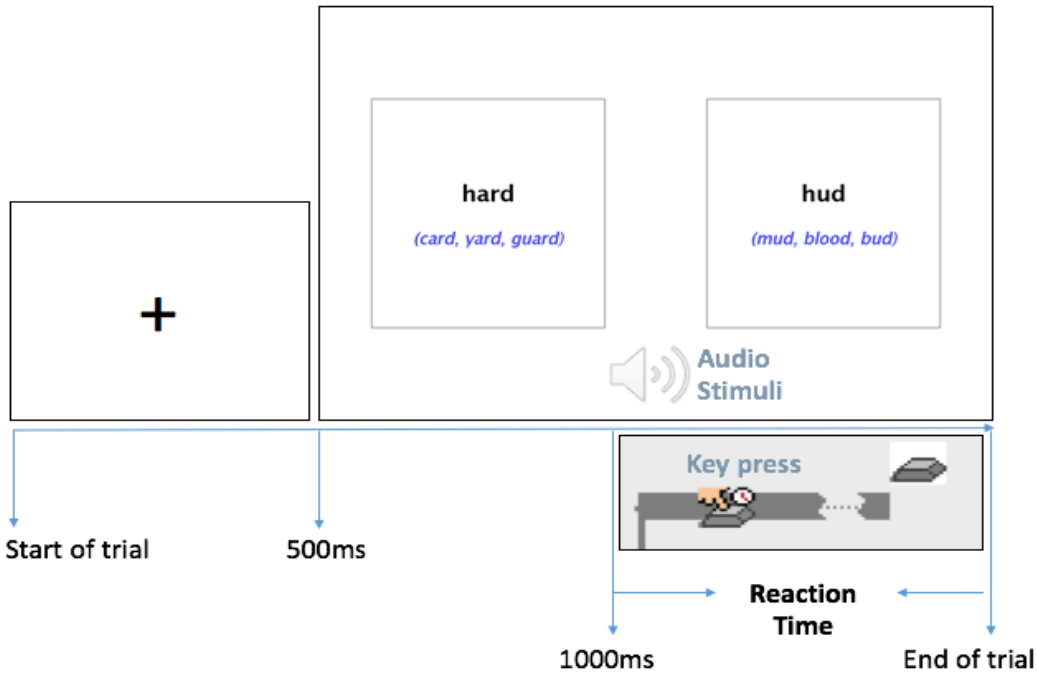
approximately 15 minutes. After finishing the 88 actual trials, listeners were asked to rest for at least two minutes before they could move to the next block.

Although listeners had heard the natural target words nine times each at the familiarisation and practice stage, it was possible for them (especially L2 listeners) to forget the "vowel-word" connection during the actual test and provided responses contradictory to their perception. To decrease this possibility, from the familiarisation stage till the actual test, rhymes of the target words were presented to remind the listener of the correct pronunciation. Figure 3.3.1 illustrates how the targets words and rhymes were displayed on the screen and the timeline of different events on an actual trial. Table 3.3.2 presents all the rhymes for the target words.

Table 3.3.2 Rhymes for the target words

	Voiced Context /hVd/		Voiceless Context /hVt/	
Vowel	**Target word**	**Rhymes**	**Target word**	**Rhymes**
/ɐː/	hard	card, yard, guard	heart	cart, dart, part
/ɐ/	hud	mud, blood, bud	hut	cut, nut, but
/oː/	horde	lord, board, cord	hort	short, port, court
/ɔ/	hod	nod, god, odd	hot	knot, lot, pot

Figure 3.3.1 Presentation of an actual trial



The target word's displayed position (i.e. at left or right side on the screen) was counterbalanced across listeners to reduce the effect of handedness, that is right-handed

participants may provide more right-sided answers. In this case, a right-handed listener may show a tendency to choose the word displayed at the right side on the screen. However, for the same listener, the response choices were displayed in a constant way: target words representing short vowels, i.e. hud, hut, hod, hot, were always placed at one side, and target words of long vowels, i.e. hard, heart, horde, hod, were always placed at the other side. This is because listeners were tuned to the correspondence between the target words and the key "a" or "l" from the familiarisation stage, ensuring that they could respond as quickly as possible during the actual trials. Switching the order of response choices within the same block could result in an undesirable slower reaction time. There may also be tuning between the long/short-vowel choice and the left/right key as the two target words in the same block were always contrasted in vowel duration. Thus, changing the displayed position of long-vowel and short-vowel target words across blocks could result in confusion.

The four blocks' presenting sequence was also counterbalanced across listeners, for the purpose to reduce the previous block's effect on the following block. To give an example of the block's effect, listeners who undertook the /ɐ/- /ɐː/ blocks first may form the impression that the temporal cue was quite reliable for the task, as /ɐ/ and /ɐː/ were mainly contrasted in duration. Thus when they move to the /ɔ/-/oː/ blocks, they may show a preference of durational cue over the spectral cue due to inertia. The blocks were counterbalanced as follows. The sequence of /ɐ/- /ɐː/ blocks and /ɔ/-/oː/ blocks were counterbalanced. Half of the Mandarin and AusE listeners were presented with the /ɐ/- /ɐː/ blocks before the /ɔ/-/oː/ blocks. The other half were presented with the /ɔ/-/oː/ blocks before the /ɐ/- /ɐː/ blocks. Based on this general sequence, the order of voiced blocks (i.e. /hVd/ blocks) and voiceless blocks (i.e. /hVt/ blocks) were counterbalanced as well. Half of the listeners were presented with voiced blocks before voiceless blocks. The other half were presented with the blocks in a reversed order. As a result, the blocks were displayed in four different orders. Along with the counterbalancing of target words' displayed position, eight presenting orders were created, which is summarized in Table 3.3.3. Note that two blocks examining the same vowel contrast were separated intentionally to add more variation to the task.

Table 3.3.3 The presenting order of blocks and target words

	Block 1	Block 2	Block 3	Block 4
Version	**Left - Right**	**Left - Right**	**Left - Right**	**Left - Right**
1	hard - hud	horde - hod	heart - hut	hort - hot
2	hud - hard	hod - horde	hut - heart	hot - hort
3	horde - hod	hard - hud	hort - hot	heart - hut
4	hod - horde	hud - hard	hot - hort	hut - heart
5	heart - hut	hort - hot	hard - hud	horde - hod
6	hut - heart	hot - hort	hud - hard	hod - horde
7	hort - hot	heart - hut	horde - hod	hard - hud
8	hot - hort	hut - heart	hod - horde	hud - hard

The experiment recorded listeners' categorisation responses and reaction time (RT). The reaction time measured how long it took for a listener to respond since the onset of the stimuli was played. The reason for measuring RT from the onset rather than from the vowel was that the onset /h/ could provide articulatory information on the following vowel. /h/ is glottal fricative consonant, which does not have a lingual articulation (Ladefoged & Johnson, 2014). When it precedes a vowel, it typically assumes the articulatory gesture of the vowel, which is a case of anticipatory coarticulation (Cox, 2012). Thus, listeners could detect the vowel's information during the perception of the onset /h/ and reacted from then. This is also the reason for equalizing the stimuli's onset duration. Stimuli generated from different natural tokens had different onset duration. In other words, stimuli containing different vowels or the same vowel in different context did not provide equal amount of vowel information during the onset. Thus, the onset duration was equalized (see §3.2.2) to ensure the effect of vowel duration manipulation on the categorical responses and RT was comparable across vowels and contexts.

After the experiment, participants were also asked to fill a language background questionnaire. The information collected has been summarized in the first section (§3.1 Participants) of this chapter. The next chapter will discuss the analysis of the categorisation responses and RT data from the experiment.

Chapter 4 Data analysis and Results

This chapter reports the analysis of the categorization and RT data from the experiment to answer the research questions raised at the end of the literature review (Chapter 2) and repeated here for convenience:

Research Question 1: Do Mandarin and AusE listeners use a duration-based category to perceive the AusE vowel contrast /ɐ/ - /ɐː/? If the answer is positive, are the category boundaries similar or different between Mandarin and AusE listeners in terms of location and steepness?

Research Question 2: Do Mandarin and AusE listeners use a duration-based category boundary to perceive the AusE vowel contrast /ɔ/ - /oː/? If the answer is positive, are the category boundaries similar or different between Mandarin and AusE listeners in terms of location and steepness?

Research Question 3: If Mandarin and AusE listeners show duration-based categories, will coda voicing influence the location of their category boundaries?

4.1 Data analysis

This section describes the data analysis procedures. Prior to analysing, data were cleaned from errors and trimmed for outliers, the standards and procedures of which are reported in §4.1.1. This is followed in §4.1.2 by a descriptive analysis which will shed light on the three research questions. §4.1.3 presents the statistical mixed effects models to answer the three research questions. All analyses were conducted using the R program (R Core Team, 2016).

4.1.1 Data clean and trim

The raw data set included 22560 observations in total, of which 1440 were from the practice trials (8 natural tokens × 3 repetitions × 60 participants = 1440) and 21120 were from the actual identification task (88 stimuli × 4 repetitions × 60 participants = 21120).

Data from the practice trials were inspected to examine each participant's accuracy in identifying the original tokens. Three participants in the AusE group did not achieve 100% accuracy; the lowest accuracy in the AusE group was 87.5%. Thirteen participants in the Mandarin group made errors; the lowest accuracy in this group was 83%. No one in either AusE or Mandarin group made errors in all three tests of each token. See Appendix C for the summary of the practice data. No participant was excluded based on the results of the

practice trials, as the test was intended for familiarisation purpose only in which the participants were told to take the time to be familiarised, rather than respond as soon as possible. Furthermore, the errors could result for various reasons, e.g., lapse of concentration, problem of mapping sound with written form, or difficulty in discriminating the sound pair. Excluding L2 participants on the basis of such errors would impede us from examining the research questions, as it is inherent to L2 learning to have problems mapping the sound with written form or find it more difficult to discriminate some vowel pairs than native AusE listeners were expected to.

Data from the actual identification task were cleaned and trimmed by examining the RT distributions. Two observations were removed as errors because the corresponding RT was far from the common value. One was from an AusE participant with an RT of 51ms, which is much shorter than the lowest latency required for human brain to process auditory stimuli and prepare a response (i.e., 200ms, following Baayen & Milin, 2015). The other was from a Mandarin participant with an RT of 19699ms. There was no other observation with an RT over 10000ms and the experiment log showed that the participant indeed paused once in the middle of the test.

828 observations were removed when their RTs were trimmed as outliers. The RT was trimmed per subject and per stimulus, using *Mean – 1.458SD* as the lower criterion and *Mean+ 1.458SD* as the upper criterion. In other words, RTs 1.458 standard deviations above and 1.458 standard deviations below the mean RT for each participant responding to each stimulus was removed. The trimming was conducted separately for each participant because age and language background could influence the participants' RT (Baayen & Milin, 2015; Woods, Wyma, Yund, Herron, & Reed, 2015). Trimming observations based on a group mean and standard deviation could result in the removal of almost all observations for some participants. Similar considerations applied to the "per stimulus" method. Stimuli with varying vowel duration and vowel quality result in varied RT, which reflects the listeners' perceptual strategy and difficulty and is one of the main concerns for the current study. Therefore, it was not appropriate to place a standard RT threshold for exclusion across the board. The standard deviation to set the cut-off point in the trimming was set to 1.458 based on the sample size of the per-participant and per-stimulus RTs, following Selst & Jolicoeur (1994). Each stimulus was repeated four times for each participant, thus the RT sample size was four. Selst & Jolicoeur (1994) suggested to use 1.458SD as the trimming criterion for

this sample size[6]. 47% of the removed observations (i.e. 387 observations) were from the Mandarin group and 53% from the AusE group (i.e. 423 observations). Figure 4.1 and 4.2 below show the two groups' logged RTs before and after trimming. The large number of apparent outliers after trimming (Figure 4.2) show that there was big variation in between-subject and between-stimulus performance, even within the same language group. The cleaned and trimmed data set had 20292 observations of 60 participants' categorization and RT performance in the actual identification test.

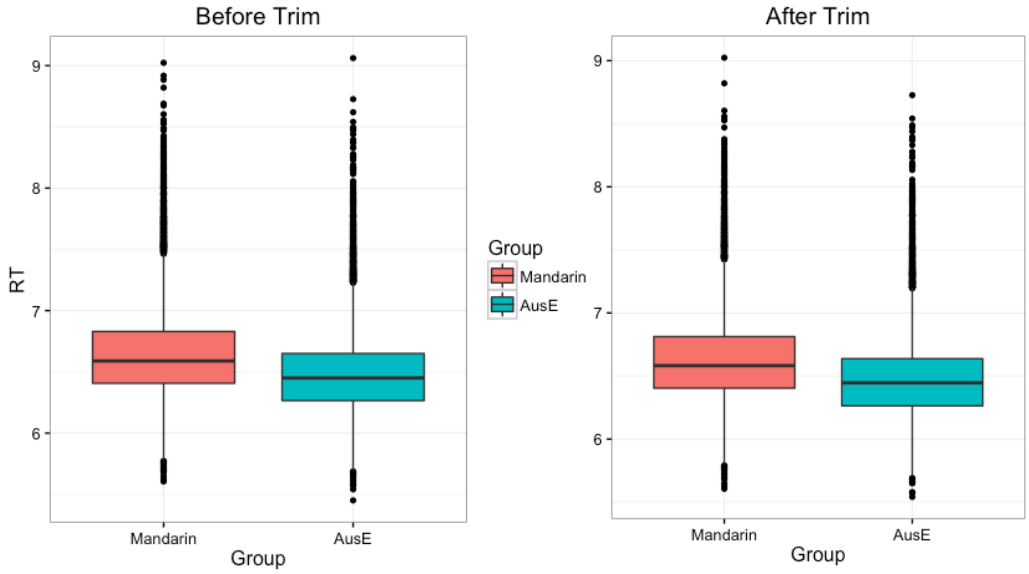


Figure 4.1 RTs before trimming

Figure 4.2 RTs after trimming

4.1.2 Descriptive analysis

Analysis of the categorization data

For each vowel pair in each context, the percentage of the long vowel answers was calculated for each vowel source (one of the eight original tokens), per duration step and per group. For instance, for the *hard-hud* block, the percentage of *hard* answers to either *hard*-based stimuli or *hud*-based stimuli at each of the 11 duration steps was calculated for both the Mandarin and the AusE group. For the *horde-hod* block, the percentage of *horde* answers was calculated. The calculation formula is as below. The denominator also equalled to the total observations of each vowel source at each duration step from each group, as it was a two-forced-choice identification task.

$$Percentage\ of\ long\ vowel\ answers = \frac{Number\ of\ long\ vowel\ answers}{Number\ of\ (long\ vowel\ answers + short\ vowel\ answers)}$$

[6] When the sample size increases, the suggested SD criterion increases.

A line graph with the duration steps on the X-axis and the percentage of long vowel answers on the Y-axis was drawn for this vowel pair. The graph shows how the percentage of long vowel answers changes as the vowel duration of the stimuli increases. Three features on the graph are important for analysis: line slope, 50% crossover point and percentages at the two endpoints of the duration continuum.

Line slope indicates whether the group's vowel categorization was influenced by duration. If the line is flat, the categorization was not influenced by vowel duration. The group mainly or totally relied on the source vowel's spectral properties to categorize the vowel. If the line has an overall increasing (i.e. positive slope) or decreasing trend (i.e. negative slope), the group's categorization was influenced by duration. However, the line was expected to have an increasing trend, as the percentage of long vowel answers increased along the duration continuum. If duration influences listeners' perception, they can rely on duration either partly, or fully. If a group relied partly on duration, the group was influenced by duration, but categorisation decisions were still mainly based on the source vowel's spectral properties. If listeners relied fully on duration, their perception was not only influenced by duration but also primarily relied on duration and not on the source vowels' spectral properties to make the categorisation.

A dynamic line without 50% crossover indicates that the group categorize the vowel based on the source vowel's spectral properties and was influenced by duration. The group tended to categorize the vowel always as long vowel (the line is above 50% all the time) or the short vowel (the line is below 50% all the time). A line with 50% crossover indicates that the group categorized the vowel based on duration. The below 50% portion shows that the group provided less than 50% long vowel answers and hence more than 50% short vowel answers at shorter duration steps. In other words, the group tended to categorize the vowel with shorter duration as the short vowel. In contrast, the above 50% portion indicates that the group tended to categorize the vowel at longer duration steps as the long vowel. There is a categorization change due to the vowel duration variation. Thus the group is reported to categorize the vowel based on vowel duration. The category boundary between short and long vowels is the duration point corresponding to the 50% crossover. The duration point could be between two duration steps, or at one duration step. The steepness of the boundary can be decided by "the rate of change from one category to the other in the direction perpendicular to the orientation of the boundary" (Morrison, 2007, p. 15). In this case, the rate of change was determined as the percentage increase within two duration steps that were

closest to the category boundary. The steepness of the boundary (rate of change) was computed in the following formula:

$$Steepness = \frac{Percentage\ at\ Step\ (N+1) - Percentage\ at\ Step\ N}{1}$$

The way to decide Step (N+1) and Step N:

When the boundary is located between two duration steps, Step N and Step (N+1) are the two duration steps. The boundary corresponds to 50% crossover, thus the Percentage at Step N < 50%, and the Percentage at Step (N+1) > 50%.

When the boundary is at a certain duration step, Step (N+1) is that duration step and Step N is the step before Step (N+1). The Percentage at Step N < 50% and the Percentage at Step (N+1) = 50%.

The percentages at the endpoints of the duration continuum indicates the group's certainty about its categorization. The closer the percentages are to 50%, the less certain the group was. For instance, when the line has an overall increasing trend with 50% crossover, the group relied on vowel duration to make the categorization. If the percentage at the start was close to 0% and the percentage at the end was close to 100%, the group was quite certain about duration-based categorization. If either the percentage is close to 50%, the group was less certain about its categorization. The uncertainty could result from individual variation: some participants in the group made categorisation based on spectral properties instead of duration.

Analysis of the RT data

For each vowel pair in each context, the mean logged RT was calculated for per vowel source, per duration step and per group. A line graph was drawn with the duration steps on the X-axis and mean logged RT on the Y-axis. The graph shows how the mean RT of a group changes with a varying vowel duration. It provides further support to the observations from the categorization data.

If the group shows duration-based categorization on the graph of the categorization data, the corresponding RT line may have a peak near the duration point representing the category boundary, which shows the group had a higher RT near the boundary and lower RT within each duration category. This is because the stimuli near the boundary have greater ambiguity and represent the "short" or "long" category to a lesser extent than those further away from the boundary (Massaro, 1987). Thus processing the stimuli near the category

boundary takes longer time (Schneider, Dogil, & Möbius, 2011). The point of the RT line corresponding to the longer duration values is also expected to be higher than mean RT corresponding to the shorter duration values due to a general effect of vowel duration on response latencies: longer vowel durations may lead to higher RT.

If the group shows a categorization based on a combination of the source vowel's spectral quality and an influence of the vowel duration, the RT line should either increase slowly or decrease slowly. RT increases along the duration continuum when the source vowel is intrinsically short and decrease along the duration continuum when the source vowel is intrinsically long. As mentioned above, a short source vowel with a long duration and a long source vowel with a short duration are ambiguous for listeners who are influenced by both spectral and durational information. The uncertainty resulting from such ambiguity could lead to higher RT (Pisoni & Tash, 1974; Schneider et al., 2011).

If the group shows a spectral categorisation without any duration influence, the RT line may be flat or may increase slowly along the duration continuum. This depends on the listeners' strategy. Recall that the RT was measured from the onset /h/ in the stimuli. As /h/ carries the spectral characteristics of the following vowel, listeners who relied on spectral properties only could provide answers without hearing the actual vowel. Their RTs would be similar across all the duration steps. However, some listeners who have a more cautious character may choose to respond after the whole stimulus was played. The RTs would then increase slowly along the duration continuum.

Summary

A short summary of how the descriptive analysis can shed light on the three research questions is given here. ***Research Question 1 &2*:** on the line graph of the categorization data, the line shape, 50% crossover, and percentages at the two endpoints were examined. On the line graph of the RT data, the existence of peak and overall trend were examined. An increasing categorization line with 50% crossover and an RT line with a peak indicate a duration-based categorization. The location of the category boundary is the duration point corresponding to the 50% crossover and to the RT peak. The steepness of the boundary is the rate of percentage change near the boundary. ***Research Question 3:*** categorization graphs for the same vowel pair in different coda contexts were compared in terms of the 50% crossover point.

4.1.3 Statistical analysis

Statistical analysis was carried out to answer the questions. Mixed-effects logistic regression models were fit to the categorization data and mixed-effects linear regression

models were fit to the RT data, both using the *lme4* package in R (Bates, Mächler, Bolker, & Walker, 2015). For each vowel pair in each context, categorization responses were modelled with fixed predictors Duration (85-335ms), Source Vowel , Language (Mandarin, AusE), all interactions, and a by-participant random intercept[7]. Logged RT were modelled with fixed predictors Duration (linear and quadratic effect), Source Vowel, Language (Mandarin, AusE), all interactions, and by-participant random intercepts. The formulas for the models were as follows:

Long-Vowel Response ~ Duration × Source Vowel × Language + (1|Participant)

log(RT) ~ (Duration linear + Duration quadratic) × Source Vowel × Language + (1|Participant)

Note: RT was logged; poly (Duration, 2) includes both the linear and quadratic effect of duration – the latter was included in RT models to examine the peak on the RT line.

In the model, all the variables were coded as numerical. Responses were coded as 0 = short vowel answer and 1=long vowel answer. Duration ranged from -5 to 5 with -5 representing the endpoint of the duration continuum 85ms and 5 representing the other endpoint 335ms. Source Vowel was coded as -1 = source tokens with short vowels, i.e. *hud, hod, hut, hot*, and 1 = source tokens with long vowels, i.e. *hard, horde, heart, hort*. Language was coded as -1 = AusE and 1 = Mandarin.

All effects from the main analyses were evaluated against an α level of 0.05. Post-hoc pairwise comparisons were carried out with an α level of 0.001. The following section will

[7] Random slopes were not specified in the model due to a convergence warning. Participants were expected to have differing slopes for the effect of duration and/or source vowel. However, when Duration/Source Vowel was added as the random factor, e.g., (1+ Duration)|Participant, the models gave convergence warnings. The *lme4* authors and maintainers admit that the strategy *lme4* uses for testing convergence may give warnings to well-behaved fits with large data sets (observations more than 1e5) because of the tradeoff between computational expense and accuracy, and they are in the process of finding the best strategy (R Core Team, 2016). One of the authors provided an alternative optimizer to double check the convergence (Bates et al., 2015. It worked for some of the models, but not all of them, especially those with Duration as the random factor (Bates et al., 2015). To keep consistency, the current study constructed models without random slopes, which may result in simplification in modelling. The researcher understands this limitation and will keep looking for better solutions to the problem.

present results for each vowel pair in each context. Results of descriptive analysis are presented first, followed by the results of statistical models.

4.2 Results

4.2.1 /ɐ/ - /ɐː/

Comparison One: hVd context

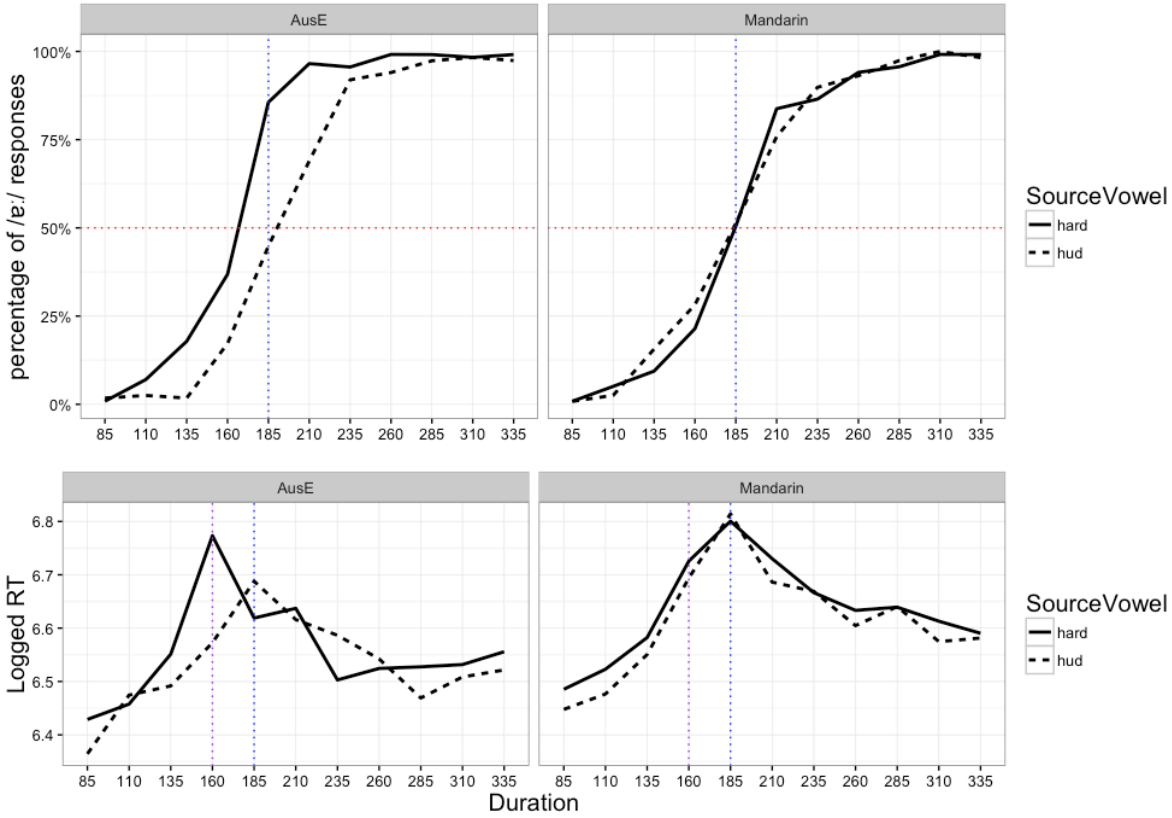


Figure 4.3 Categorization and RT line graphs for /ɐ/ - /ɐː/ in hVd context

Both the AusE and Mandarin listeners made duration-based categorization when they perceived the /ɐ/ - /ɐː/ contrast in voiced context**,** as can be seen in Figure 4.3. The percentage of long vowel responses increased consistently with vowel duration increasing and had a 50% crossover point of 160ms and 185ms. The percentages of long vowel answers at the short and long endpoints of the duration continuum were 0% and 100%, respectively, suggesting the listeners were certain about their duration-based categorization. The RT lines showed clear peaks near the 50% crossover points of their corresponding categorization line.

The category boundaries for the two groups were different in terms of location and steepness. The AusE group's category boundary for *hard*-based stimuli is near 160ms, with a steepness of 0.35 and the boundary for *hud*-based stimuli is at 185ms, with a steepness of

0.33. The Mandarin group has the same category boundary for both *hard*- and *hud*-based stimuli, which is at 185ms with a steepness of 0.30. Thus one of the AusE group's category boundaries is at an earlier duration step than all other category boundaries, suggesting that the AusE group provided more /ɐː/ answers to *hard*-based stimuli than to *hud*-based stimuli and therefore overall more /ɐː/ answers compared to the Mandarin group. In addition, the AusE group has steeper category boundaries than the Mandarin group does.

The results of the mixed-effects logistical model for categorization responses and results of the mixed-effects linear model for RT are given in Table 4.1 and 4.2, respectively. The categorization model shows a positive main effect of Duration (β=1.17, z=33.44, p<0.05) which confirms that listeners across both groups relied on duration, and a positive main effect of Source Vowel (β=0.43, z=6.49, p<0.05), which indicates that overall more "long vowel" (/ɐː/) answers were provided to *hard*-based stimuli (coded as 1) than to *hud*-based stimuli (coded as -1). The model also shows a negative main effect of Language (β= -0.28, z=-2.40, p=0.0016) indicating that the AusE group (coded as -1) provided more /ɐː/ answers than the Mandarin group (coded as 1). There are also negative effects of the Language × Duration interaction (β=-0.08, z=-2.63, p=0.008) and of the Language × Source Vowel interaction (β=-0.46, z=-6.89, p<0.05). The Language x Duration interaction confirmed that the AusE group had a steeper category boundary than the Mandarin group does.

Both interactions were further explored by subjecting the AusE and the Mandarin data to separate analyses with the predictors Duration, Source Vowel, and their interaction. Both models revealed a main effect of Duration, confirming that listeners in both groups relied on duration (AusE group: β=1.26, z=23.01, p<0.001; Mandarin group: β = 1.087, z=24.73, p<0.001). The AusE model but not the Mandarin model revealed a positive main effect of Source Vowel (AusE group: β = 0.88, z=8.29, p<0.001; Mandarin group: β = -0.03, z=-0.34, p=0.732). This indicates that only the Australian-English listeners provided more /ɐː/ answers to *hard*-based stimuli (coded as -1) than to *hud*-based stimuli (coded as 1), whereas we did not observe that the Mandarin listeners responded differently to *hard*-based and *hud*-based stimuli. A final analysis subjected the responses to the *hard*-based and *hud*-based stimuli to separate analyses with the predictors Duration, Language, and their interaction. Both models showed the expected effect of duration (*hard*-based stimuli: β = 1.26, z=22.58, p<0.001; *hud*-based stimuli: β = 1.14, z=24.03, p<0.001). The *hard*-based but not the *hud*-based analysis revealed a main effect of Language (*hard*-based stimuli: β = -0.75, z=-4.77, p<0.001; *hud*-based stimuli: β =0.16, z=1.213, p=0.225). This indicates that the Australian English listeners

provided more /ɐː/ answers to *hard*-based stimuli than the Mandarin listeners, whereas no difference in the rate of /ɐː/ answers was observed for the *hud*-based stimuli. Neither model revealed a Duration x Language interaction.

The RT model shows a positive main effect of Linear Duration (β=6.58, t=4.68, p<0.05) which indicates an overall higher RT with longer vowel duration. There is a negative main effect of Quadratic Duration (β=-4.62, z=-15.49, p<0.05) which captures the RT peak. A positive main effect of Source Vowel (β=0.011, z=2.644, p=0.008) showing higher *RT* for *hard*-based stimuli compared to *hud*-based stimuli. The Language × Quadratic Duration interaction shows an negative main effect (β=-0.062, z=-2.064, p=0.039). This negative interaction shows that the RT peak is steeper in the Mandarin English group (coded as 1) than in the Australian English group (coded as -1).

Table 4.1 Mixed-effects logistic regression model for categorization responses in *hud-hard* block

Predictor	*ß*	*SE*	*z*	*p*
Intercept	1.36920	0.11891	11.52	< 2e-16 ***
Duration	1.17338	0.03509	33.44	< 2e-16 ***
SourceVowel	0.42919	0.06616	6.49	8.76e-11 ***
Language	-0.28484	0.11847	-2.40	0.01620 *
Language× Duration	-0.08920	0.03388	-2.63	0.00847 **
Language× SourceVowel	-0.45599	0.06618	-6.89	5.59e-12 ***
Duration× SourceVowel	0.05599	0.03214	1.74	0.08153
Language×Duration× SourceVowel	-0.04094	0.03214	-1.27	0.20279

Table 4.2 Mixed-effects linear regression model for RT in *hud-hard* block

Predictor	*ß*	*SE*	*t*	*p*
Intercept	6.583e+00	2.104e-02	312.875	< 2e-16 ***
LinearDuration	1.395e+00	2.981e-01	4.681	2.93e-06 ***
QuadraticDuration	-4.617e+00	2.981e-01	-15.489	< 2e-16 ***
SourceVowel	1.106e-02	4.184e-03	2.644	0.00823 **
Language	4.004e-02	2.104e-02	1.903	0.06186
Language×LinearDuration	4.362e-01	2.981e-01	1.464	0.14338
Language×QuadraticDuration	-6.153e-01	2.981e-01	-2.064	0.03905 *
Language×SourceVowel	2.645e-04	4.184e-03	0.063	0.94959
LinearDuration×SourceVowel	-3.622e-01	2.981e-01	-1.215	0.22437
QuadraticDuration×SourceVowel	4.019e-01	2.981e-01	1.348	0.17759
Language× LinearDuration× SourceVowel	1.225e-01	2.981e-01	0.411	0.68107
Language×QuadraticDuration×SourceVowel	-1.444e-01	2.981e-01	-0.484	0.62819

Comparison TWO: hVt context

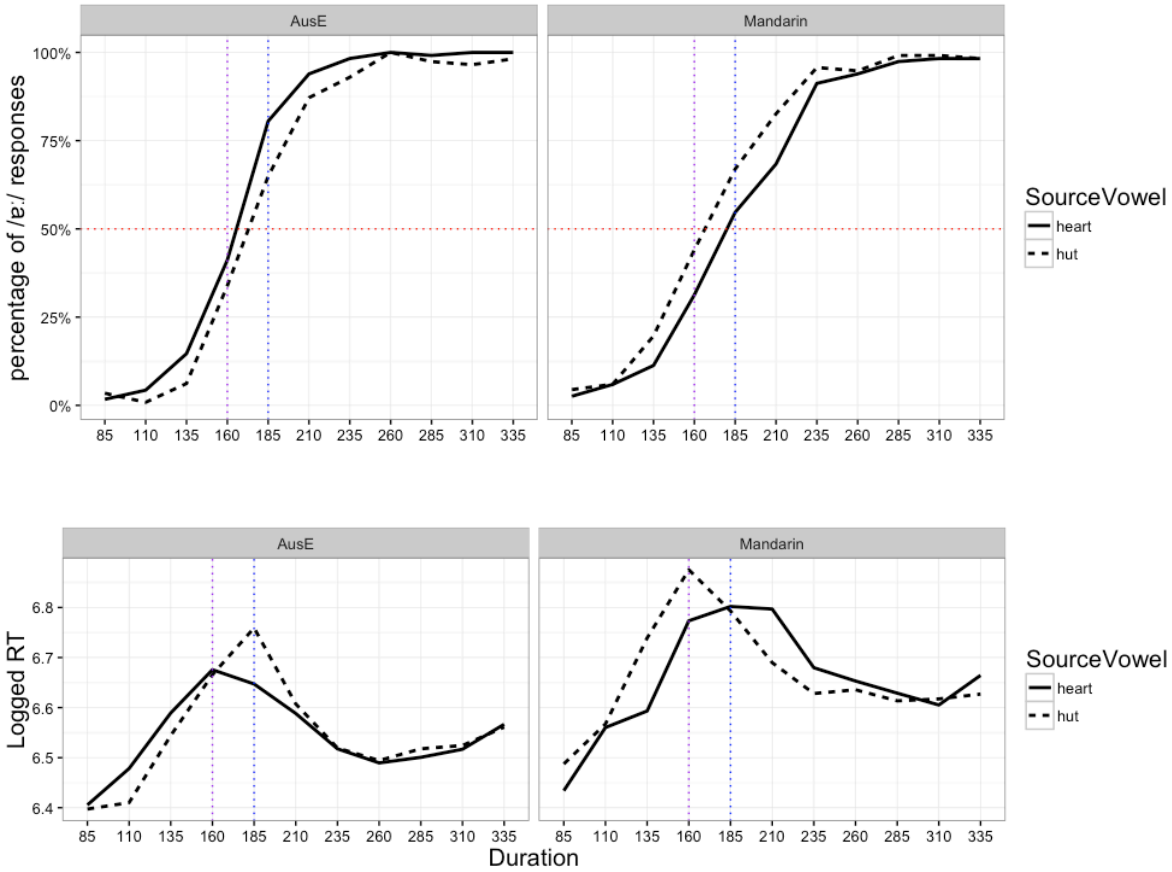


Figure 4.4 Categorization and RT line graphs for /ɐ/ - /ɐː/ in hVt context

Both the AusE and Mandarin group made duration-based categorization when they perceived the /ɐ/ - /ɐː/ contrast in voiceless context as well**,** as can be seen in Figure 4. The four category boundaries fall between 160ms and 185ms. However, the AusE group had a boundary at lower duration values for *heart*-based stimuli while the Mandarin group had a boundary at lower duration values for *hut*-based stimuli. The AusE group's boundaries are also steeper than the Mandarin group's: 0.38 compared to 0.27 for both *heart*- and *hut*-based stimuli.

The results of the mixed-effects logistical model for categorization responses and results of the mixed-effects linear model for RT are given in Table 4.3 and 4.4, respectively. The categorization model shows a positive main effect of Duration (β=1.19, z=31.64, p<0.05) which confirms that listeners across both groups relied on duration. There is no effect of Source Vowel (β=0.17, z=2.27, p=0.023) which indicates similar number of "long vowel" (/ɐː/) answers were provided to *heart*-based stimuli and *hut*-based stimuli. The model also shows a negative main effect of Language (β= -0.46, z=-3.86, p < 0.05) indicating that the

AusE group (coded as -1) provided more /ɐː/ answers than the Mandarin group (coded as 1). There are also negative effects of the 2-way interactions Language × Duration (*ß*=-0.21, *z*=-5.63, *p*<0.05) and Language × Source Vowel (*ß*=-0.49, *z*=-6.52, *p*<0.05). The Language x Duration interaction confirms that the AusE group has a steeper category boundary than the Mandarin group does.

Both interactions were further explored by subjecting the AusE and the Mandarin data to separate analyses with the predictors Duration, Source Vowel, and their interaction. Both models revealed a main effect of Duration, confirming that listeners in both groups relied on duration (AusE group: *ß*=1.42, *z*=21.63, *p*<0.001; Mandarin group: *ß* = 0.97, *z*=25.12, *p*<0.001). The AusE model revealed a positive main effect of Source Vowel (*ß* = 0.67, *z*=5.27, *p*<0.001). The Mandarin model revealed a negative main effect of Source Vowel (*ß* = -0.32, *z*=-3.88, *p*=0.001). This indicates that the Australian English listeners provided more /ɐː/ answers to *heart*-based stimuli (coded as 1) than to *hut*-based stimuli (coded as -1), whereas Mandarin listeners provided more /ɐː/ answers to *hut*-based stimuli (coded as -1) than to *heart*-based stimuli (coded as 1). A final analysis subjects the responses to the *hard*-based and *hud*-based stimuli to separate analyses with the predictors Duration, Language, and their interaction. Both models showed the expected effect of duration (*heart*-based stimuli: *ß* = 1.29, *z*=21.03, *p*<0.001; *hut*-based stimuli: *ß* = 1.10, *z*=23.70, *p*<0.001). The *heart*-based but not the *hut*-based model revealed a main effect of Language (*heart*-based stimuli: *ß* = -0.96, *z*=-6.00, *p*<0.001; *hut*-based stimuli: *ß* =0.04, *z*=0.30, *p=0.77*). This indicates that the Australian English listeners provided more /ɐː/ answers to *heart*-based stimuli than the Mandarin listeners, whereas no difference in the rate of /ɐː/ answers was observed for the *hut*-based stimuli. Neither model revealed a Duration x Language interaction.

The RT model shows a positive main effect of Linear Duration (*ß*=0.88, *t*=3.10, *p*<0.05) which indicates an overall higher RT with longer vowel duration. There is a negative main effect of Quadratic Duration (*ß*=-0.43, *z*=-15.06, *p*<0.05) which captures the RT peak. There is no effect of Source Vowel (*ß*=-0.003, *z*=-0.68, *p*=0.50) showing that no consistent difference in *RT* was found between *heart*-based and *hut*-based stimuli. The Language × Quadratic Duration interaction is significantly negative (*ß*=-0.008, *z*=-2.963, *p*<0.005), suggesting that the RT peak is steeper in the Mandarin English group (coded as 1) than in the Australian English group (coded as -1).

Table 4.3 Mixed-effects logistic regression model for categorization responses in *hut-heart* block

Predictor	*ß*	*SE*	*z*	*p*
Intercept	1.79479	0.12082	14.86	< 2e-16 ***
Duration	1.19297	1.19297	31.64	< 2e-16 ***
SourceVowel	0.16988	0.07486	2.27	0.023254
Language	-0.46415	0.12010	-3.86	0.000111 ***
Language× Duration	-0.20595	0.03660	-5.63	1.83e-08 ***
Language× SourceVowel	-0.48862	0.07494	-6.52	7.03e-11 ***
Duration× SourceVowel	0.08425	0.03457	2.44	0.014788
Language×Duration× SourceVowel	-0.10620	0.03457	-3.07	0.002125 **

Table 4.4 Mixed-effects linear regression model for RT in *hut-heart* block

Predictor	*ß*	*SE*	*t*	*p*
Intercept	6.601e+00	2.048e-02	322.312	< 2e-16 ***
LinearDuration	8.832e-01	2.856e-01	3.093	0.00200 **
QuadraticDuration	-4.301e+00	2.856e-01	-15.060	< 2e-16 ***
SourceVowel	-2.723e-03	4.004e-03	-0.680	0.49655
Language	5.677e-02	2.048e-02	2.772	0.00741**
Language×LinearDuration	7.491e-02	2.856e-01	0.262	0.79309
Language×QuadraticDuration	-8.463e-01	2.856e-01	-2.963	0.00306 **
Language×SourceVowel	-9.931e-04	4.004e-03	-0.248	0.80413
LinearDuration×SourceVowel	3.852e-01	2.856e-01	1.349	0.17747
QuadraticDuration×SourceVowel	9.784e-02	2.856e-01	0.343	0.73192
Language× LinearDuration× SourceVowel	8.357e-01	2.856e-01	2.926	0.00345 **
Language×QuadraticDuration×SourceVowel	-5.514e-01	2.856e-01	-1.931	0.05354

4.2.2 /ɔ/ - /oː/

Comparison One: hVd context

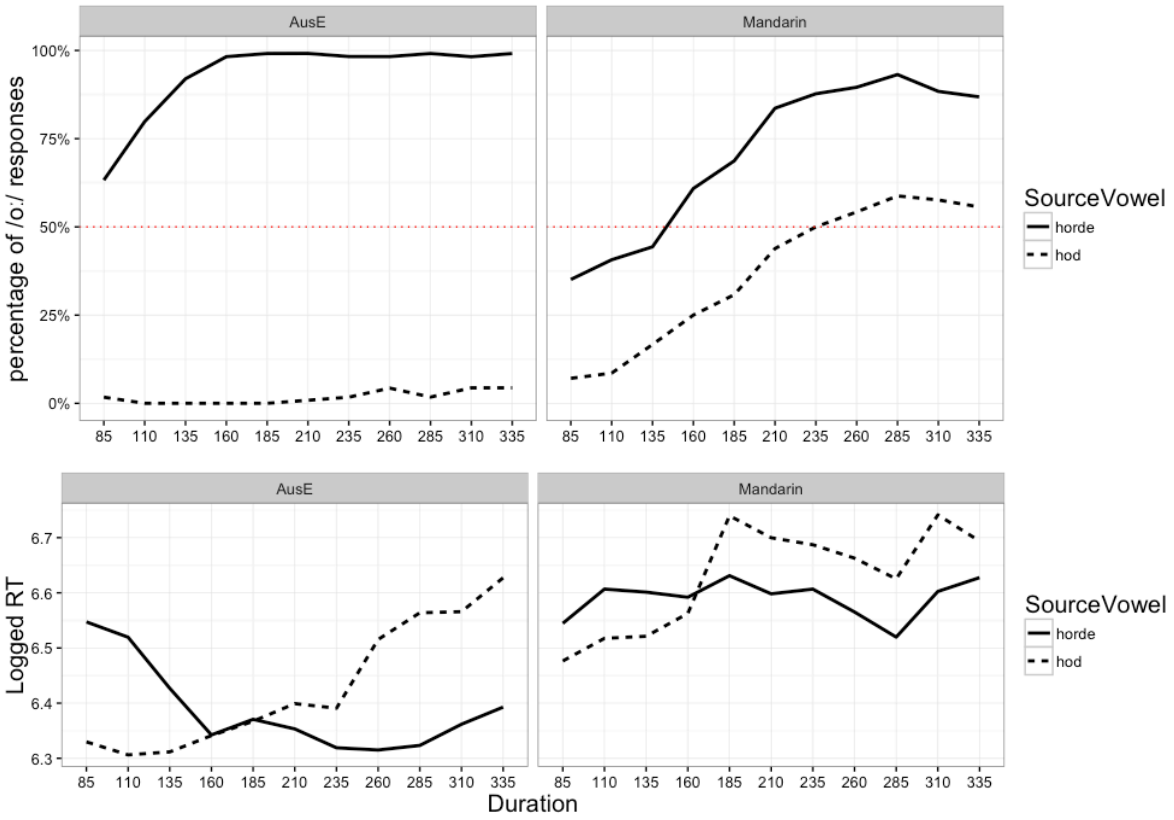


Figure 4.5 Categorization and RT line graphs for /ɔ/ - /oː/ in hVd context

The AusE and Mandarin relied to different extents on duration in their perception of the /ɔ/ - /oː/ contrast in hVd context, as can be seen in Figure 4.5. The AusE listeners always gave more than 50% /oː/ responses to all *horde*-based stimuli and almost 100% /ɔ/ responses to all *hod*-based stimuli. This means that duration has little influence on their categorization. Their consistency in responding /oː/ to *horde*-based stimuli increased with duration, and reached 100% certainty for stimuli longer than 160ms. Their increasing certainty for *horde*-based stimuli with a longer duration is confirmed by the decreasing trend of RT line for *horde*-stimuli, which reaches consistently low RT for stimuli longer than 160ms. For *hod*-based stimuli, the group was overall very consistent in the proportion of /oː / responses, and only provided a little more /oː/ answers to stimuli with a duration over 285ms. The increasing trend of the RT line for *hod*-based stimuli reflects that listeners' processing is affected by the durational information. Most likely, with duration increasing, listeners get more uncertain about categorizing *hod*-based stimuli as short /ɔ/.

The Mandarin listeners were influenced to a larger extent by duration than the AusE group. They gave less than 50% /oː/ responses to the shortest stimuli, and less than 100% /oː/ responses to the longest stimuli, meaning that with duration increasing, their responses change from primarily /ɔ/ to primarily /oː/ responses. The Mandarin listeners were also influenced by the spectral properties of the stimuli: the categorization line for *horde*-based stimuli is consistently higher than the categorization line for *hod*-based stimuli, meaning that Mandarin listeners gave more /oː/ responses to *horde*-based stimuli independently of the stimulus duration. This implies that they can detect the source vowel information from the stimuli. In fact, the categorization line for *horde*-based stimuli starts just below 50% for the shortest stimuli, while the line for *hod*-stimuli ends just above 50% for the longest stimuli. This means that the Mandarin listeners, as a group, are not certain about their responses to these stimuli with conflicting spectral and durational cues.

The RT lines for the Mandarin group show similar trend as the RT lines for the AusE group, but the Mandarin group's overall RT is much higher than the AusE group, confirming that they were more uncertain about their responses or required more processing time.

The results of the model for categorization responses and results of the model for linear for RT are given in table 4.5 and 4.6, respectively. The categorization model showed positive effects for all three main effects. The effect of Duration (β=0.39, z=13.55, p<0.05), confirmed that listeners responded more /oː/ to longer stimuli. The effect of Source Vowel (β=2.59, z=24.03, p<0.05), confirmed that listeners responded more /oː/ to *horde*-based stimuli (coded as 1) than to *hod*-based stimuli (coded as -1). The effect of Language (β=0.25, z=2.091, *p=0.037)* suggested that Mandarin listeners responded more /oː/ than AusE listeners. The negative effect of the Language × Duration interaction (β=-0.07, z=-2.33, *p=0.020*) shows that the Mandarin group relied more on Duration than the AusE group. The negative effect of the Language × Source Vowel interaction (β=-1.70, z=-15.83, *p<0.05*) shows the AusE group was significantly more influenced by the characteristics of the Source vowel than the Mandarin group. The positive effect of the Duration × Source Vowel interaction (β=0.10, z=0.03, *p<0.05*) shows that the perception of *horde*-based stimuli was significantly more influenced by duration, compared to *hod*-based stimuli.

The RT model shows a positive main effect of Linear Duration (β=2.26, t=7.30, p<0.05), Quadratic Duration (β=0.84, t=2.73, p<0.05) and Language (β=0.10, z=4.51, *p<0.05)*. The positive main effect of Quadratic Duration (β=0.84, t=2.73, p<0.05) suggests that averaged over the language groups and source vowels there was u-shaped relationship

between RT and duration. There is also a negative main effect of Source Vowel (*ß*=-0.019, *t*= -4.575, *p*<0.05), indicating that participants responded with longer RT to *hod*-based compared to *horde*-based stimuli.

Table 4.5 Mixed-effects logistic regression model for categorization responses in *hod-horde* block

Predictor	*ß*	*SE*	*z*	*p*
Intercept	0.005097	0.119906	0.043	0.9661
Duration	0.388574	0.028687	13.545	<2e-16 ***
SourceVowel	2.591236	0.107805	24.036	<2e-16 ***
Language	0.250626	0.119881	2.091	0.0366 *
Language× Duration	-0.066689	0.028607	-2.331	0.0197 *
Language× SourceVowel	-1.693279	0.106976	-15.829	<2e-16 ***
Duration× SourceVowel	0.103464	0.028621	3.615	0.0003 ***
Language×Duration× SourceVowel	-0.067123	0.028614	-2.346	0.0190 *

Table 4.6 Mixed-effects linear regression model for RT in hod-horde block

Predictor	*ß*	*SE*	*t*	*p*
Intercept	6.509e+00	2.241e-02	290.504	< 2e-16 ***
LinearDuration	2.253e+00	3.086e-01	7.302	3.29e-13 ***
QuadraticDuration	8.442e-01	3.086e-01	2.735	0.00625 **
SourceVowel	-1.985e-02	4.338e-03	-4.575	4.87e-06 ***
Language	1.011e-01	2.241e-02	4.512	3.05e-05 ***
Language×LinearDuration	3.749e-01	3.086e-01	1.215	0.22453
Language×QuadraticDuration	-2.090e+00	3.086e-01	-6.773	1.41e-11 ***
Language×SourceVowel	8.061e-04	4.338e-03	0.186	0.85259
LinearDuration×SourceVowel	-3.956e+00	3.086e-01	-12.819	< 2e-16 ***
QuadraticDuration×SourceVowel	9.073e-01	3.086e-01	2.940	0.00330 **
Language× LinearDuration× SourceVowel	1.587e+00	3.086e-01	5.142	2.83e-07 ***
Language×QuadraticDuration×SourceVowel	7.665e-02	3.086e-01	0.248	0.80384

Comparison TWO: hVt context

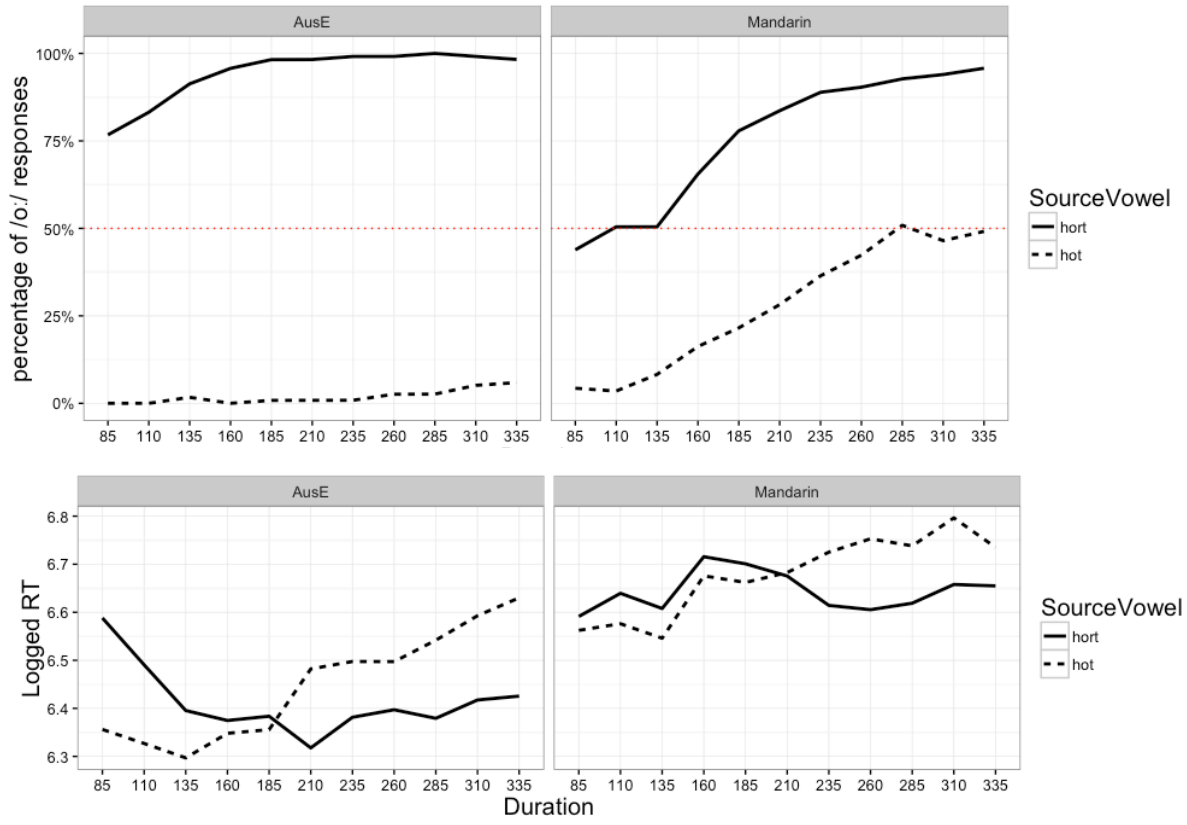


Figure 4.6 Categorization and RT line graphs for /ɔ/ - /oː/ in hVt context

The AusE and Mandarin relied to different extents on duration in their perception of the /ɔ/ - /oː/ contrast in hVt context, as can be seen in Figure 4.6. The AusE listeners always gave more than 50% /oː/ responses to all *hort*-based stimuli and almost 100% /ɔ/ responses to all *hot*-stimuli. This means that duration has little influence on their categorization. Their consistency in responding /oː/ to *hort*-based stimuli increased with duration, and reached 100% certainty after 185ms. Their increasing certainty for horde-based stimuli with a longer duration is confirmed by the decreasing trend of RT line for *horde*-stimuli, which reaches consistently low RT at 185ms and beyond. For *hot*-based stimuli, the group was overall very consistent in the proportion of /ɔ/ responses, and only provided a little more /oː/ answers to stimuli with a duration over 235ms. The increasing trend of the RT line for *hot*-based stimuli reflects that their processing is affected by the durational information. Most likely, with duration increasing, listeners get more uncertain about categorizing *hot*-based stimuli as short /ɔ/.

The Mandarin listeners were influenced to a larger extent by duration than the AusE group. They gave near 50% /o:/ responses to the shortest *hort*-based stimuli which increased to less than 100% /o:/ responses to the longest *hort*-based stimuli. The Mandarin listeners were also influenced by the spectral properties of the stimuli: the categorization line for *hort*-based stimuli is consistently higher than the categorization line for hot-based stimuli, meaning that they can detect the source vowel information from the stimuli. In fact, the categorization line for *hort*-based stimuli starts just below 50% for the shortest stimuli, while the line for *hot*-stimuli ends just near 50% for the longest stimuli. This means that they are not certain about their responses. The RT lines for the Mandarin group show similar trend as the RT lines for the AusE group, but the Mandarin group's overall RT is much higher than the AusE group, confirming that they were more uncertain about their responses or required more processing time.

The results of the model for categorization responses and results of the model for linear for RT are given in table 4.6 and 4.7, respectively. The categorization model showed two main effects. The positive effect of Duration (β=0.40, z=13.38, p<0.05) confirmed that listeners responded more /o:/ for longer stimuli. The positive effect of Source Vowel (β=2.87, z=25.30, p<0.05), confirmed that listeners responded more /o:/ to *hort*-based stimuli (coded as 1) than to *hot*-based stimuli (coded as -1). There is no effect of Language (β=0.25, z=1.841, *p=0.07),* suggesting that there may be no differences between the overall proportion of /o:/ answers across the two groups. The negative effect of the Language × Source Vowel interaction (β=-1.51, z=-13.46, *p<0.05*) shows the AusE group was significantly more influenced by Source vowel than the Mandarin group was. The positive effect of the Duration × Source Vowel interaction (β=0.05, z=2.03, *p=0.04*) shows that the perception of horde-based stimuli was significantly more influenced by duration, compared to *hod*-based stimuli.

The RT model shows a positive main effect of Linear Duration (β=2.69, t=8.70, p<0.05), Quadratic Duration (β=0.76, t=2.473, p<0.05) and Language (β=0.11, z=4.607, *p<0.05)*. The positive main effect of Quadratic Duration (β=0.76, t=2.473, p<0.05) suggests that averaged over the language groups and source vowels there was u-shaped relationship between RT and duration. This main effect was qualified by the increasing RT line for *hot*-based stimuli and declining RT line for *hort*-based stimuli. There is also a positive main effect of Source Vowel (β=-0.02, t= -4.099, p<0.05), indicating that *hot*-based stimuli caused lower RT on average compared to *hort*-based stimuli.

Table 4.7 Mixed-effects logistic regression model for categorization responses in *hot-hort* block

Predictor	*ß*	*SE*	*z*	*p*
Intercept	-0.07390	0.13316	-0.555	0.5789
Duration	0.39654	0.02963	13.383	<2e-16 ***
SourceVowel	2.87225	0.11352	25.302	<2e-16 ***
Language	0.24503	0.13313	1.841	0.0657
Language× Duration	-0.04504	0.02954	-1.525	0.1274
Language× SourceVowel	-1.50915	0.11216	-13.456	<2e-16 ***
Duration× SourceVowel	0.05987	0.02953	2.027	0.0426 *
Language×Duration× SourceVowel	-0.02703	0.02954	-0.915	0.3601

Table 4.8 Mixed-effects linear regression model for RT in *hot-hort* block

Predictor	*ß*	*SE*	*t*	*p*
Intercept	6.546e+00	2.497e-02	262.178	< 2e-16 ***
LinearDuration	2.693e+00	3.097e-01	8.696	< 2e-16 ***
QuadraticDuration	7.659e-01	3.097e-01	2.473	0.0134 *
SourceVowel	1.783e-02	4.350e-03	-4.099	4.21e-05 ***
Language	1.150e-01	2.497e-02	4.607	2.18e-05 ***
Language×LinearDuration	1.187e-01	3.097e-01	0.383	0.7016
Language×QuadraticDuration	-1.765e+00	3.097e-01	-5.698	1.28e-08 ***
Language×SourceVowel	-2.270e-04	4.350e-03	-0.052	0.9584
LinearDuration×SourceVowel	-3.601e+00	3.097e-01	-11.627	< 2e-16 ***
QuadraticDuration×SourceVowel	5.701e-01	3.097e-01	1.841	0.0657
Language× LinearDuration× SourceVowel	1.217e+00	3.097e-01	3.930	8.63e-05 ***
Language×QuadraticDuration×SourceVowel	-5.582e-01	3.097e-01	-1.802	0.0716

Chapter 5 Discussion

5.1 Perception of /ɐ/-/ɐː/

5.1.1 Duration effect

The results reported in Chapter 4 demonstrate that both AusE and Mandarin listeners categorize the AusE vowel pair /ɐ/-/ɐː/ based on vowel duration. The duration-based categories they exhibit differ a little in boundary location, but have significant statistical differences in boundary steepness. In voiced context, AusE listeners' boundary between the *short (/ɐ/)* and *long (/ɐː/)* category is at a slightly earlier duration point than that of the Mandarin listeners. In the voiceless context, listeners from the two language groups show similar boundary locations. In both contexts, AusE listeners' category boundaries are steeper than Mandarin listeners'. This is shown as a significant negative effect of the interaction between language and duration in the statistical analysis. AusE listeners' steeper boundary shows that they have established more robust *short* and *long* categories than Mandarin listeners have done. The shorter RTs for the AusE group also support this finding.

The fact that Mandarin listeners have a similar perception boundary between short and long categories to native AusE listeners could support the hypothesis of the L2LP. That is L2 learners from an L1 without phonological duration contrast can create duration categories in a native-like way during the L2 learning process. The Mandarin group's fuzzy category boundary might be attributed to the Mandarin participants' varying experience with AusE (Flege & Liu, 2001) and hence varying proficiency in AusE, as their length of study in Australia ranges from 0.5 year to 7 years.

Further support for the hypothesis that Mandarin learners create new duration boundaries through L2 learning instead of transferring their duration experience with tones in the L1 to L2 could be obtained using tone identification tasks. If the Mandarin listeners' categorization of the second tone (shorter) and third tone (longer) is based on different *short* and *long* categories that they can harness when perceiving AusE /ɐ/-/ɐː/, the developmental approach, i.e. L2LP, should be a more satisfactory explanation than the L1 transfer approach for L2 learners' reliance on duration in L2 sounds perception.

5.1.2 Voicing effect

AusE and Mandarin listeners' categorization of *hud*- and *hut*- based stimuli exhibit the same voicing effect. Their category boundary shifts to an earlier duration point for *hut*-stimuli, compared to *hud*-stimuli. This means the listeners identify more vowels with shorter duration as the long vowel category in voiceless context compared to in the voiced context. A

possible explanation is that listeners expect both the short /ɐ/ and long /ɐː/ to be shorter in a voiceless context than in a voiced context, as the vowels are produced with shorter length before a voiceless coda than before a voiced coda. Thus a lower boundary in duration is used to categorize vowels in the voiceless context. This finding again suggests that the Mandarin learners resemble native AusE listeners in perceiving vowel duration. This resemblance cannot be attributed to L1 transfer, as Mandarin syllables use no coda except the alveolar nasal consonant /n/.

Of particular interest is that the voicing effect does not show in the perception of *hard*- and *heart*-based stimuli by both AusE and Mandarin listeners. Both AusE and Mandarin listeners identify *heart*-based stimuli using the same *short*-*long* boundaries they use to identify the *hard*-based stimuli. A possible explanation for the lack of voicing effect is that the source vowel /ɐː/ in these stimuli is expected to have a least degree of shortening in the voiceless context. As /ɐ/-/ɐː/ only differ in duration in Australian English, maintaining the duration contrast between the two vowels is important for AusE listeners (Felicity Cox et al., 2015). To this end, the long /ɐː/ is shortened to the least degree before a voiceless coda (Felicity Cox et al., 2015). However, this explanation implies a source vowel effect on both AusE and Mandarin listeners, namely the listeners can discriminate stimuli generated from *hard*, from the stimuli generated from *hud*. They know whether the vowel in the stimuli is /ɐː/ with varied duration or /ɐ/ with varied duration. However, the statistical analysis only finds such source vowel effect on AusE listeners in the *hard*-*hud* block, which will be discussed in the next section. The study has not found a better way to explain the voicing effect issue. The issue needs future investigation.

5.1.3 Source vowel effect

AusE listeners provide more /ɐː/ responses than Mandarin listeners do in the *hard*-*hud* block. They also provide more /ɐː/ responses to hard-based stimuli than to hud-based stimuli. In contrast, the heart-hut block does not show either of the asymmetries. It is reasonable to consider that the asymmetries may result from a lexical effect, because listeners provide /ɐː/ responses by choosing the target word *hard* against the other choice *hud* representing /ɐ/. As *hard* is a real word and *hud* is a non-word, AusE listeners may show a preference to respond with *hard*. However, if AusE listeners were influenced by the lexical effect, they should have responded to all the stimuli with a preference of *hard*, but they only provide more /ɐː/ responses to *hard*-based stimuli. In fact, the number of /ɐː/ responses AusE listeners provided to *hud*-based stimuli is similar to the number of /ɐː/ responses provided by Mandarin listeners to either *hard*- or *hud*-based stimuli. Thus AusE listeners' overall asymmetry to /ɐː/ responses

compared to the Mandarin group only results from their categorization of the *hard*-based stimuli.

A possible explanation is that the AusE listeners' categorization of /ɐ/-/ɐː/ is influenced not only by duration, but also by the source vowel. AusE listeners may be able to detect some information in the stimuli which allows them to trace back to the source vowel. The information could be the vowel's target-to-transition proportion. As mentioned in the method section, long and short vowels do not only differ in the overall vowel duration, but also in the target-to-offglide ratio. Long vowels like /ɐː/ generally have a proportionately longer stable part, i.e. target, compared to short vowels like /ɐ/. Recall that the duration manipulation in this study was conducted for the whole vowel, with the target-to-transition ratio maintained in each stimulus item. AusE listeners' categorization could be influenced by the ratio. At the start of the duration continuum, all the stimuli's vowel duration is so short that AusE listeners may not be able to detect the ratio information. Their categorization at this stage is completely influenced by duration. With duration increasing, AusE listeners may be able to tell that *hard*-based stimuli contain vowels that have comparatively longer target component than those in *hud*-based stimuli. From this point, AusE listeners tend to provide more /ɐː/ answers to the *hard*-based stimuli compared to the *hud*-based stimuli. Only when the whole vowel duration increases to a certain degree do they start to categorize the vowels in *hud*-stimuli as /ɐː/. This perceptual process should also apply to the *heart*-*hut* block. However, the statistical analysis does not show more /ɐː/ answers were provided to *heart*-based stimuli compared to *hut*-based stimuli. This is because the voicing effect influenced the AusE listeners' perception and they provide more /ɐː/ answers to *hut*-based stimuli than they provide to *hud*-based stimuli. Thus the difference in the number of /ɐː/ answers between *heart*- and *hut*-based stimuli is smaller than that between *hard*- and *hud*- based stimuli.

5.2 Perception of /ɔ/-/oː/

5.2.1 Duration effect

AusE listeners and Mandarin learners of English are influenced by duration to a different extent when they categorize the AusE pair /ɔ/-/oː/. AusE listeners make the categorization based on the spectral properties of the source vowel. Varied vowel duration only influences their certainty about the choice, which is reflected in RT. With duration increasing, AusE listeners show decreasing RT to categorize *horde*-based stimuli as /oː/ and increasing RT to categorize *hod*-based stimuli as /ɔ/. This confirms previous finding that the spectral information is the primary source of the AusE vowel identity for this pair but an

integration of spectral and durational information leads to higher identification accuracy (Watson and Harrington, 1999).

Mandarin listeners are more influenced by duration than AusE listeners are. This is shown in the statistical analysis as a positive effect of language (AusE coded as -1 and Mandarin coded as 1) and duration interaction. They are also influenced by the spectral properties of the source vowel. However, the combination of spectral and durational information makes them more uncertain about their choice than AusE listeners are. As a result, their RTs are much higher than those of AusE listeners.

A finding that can shed light on those L2 speech perception models is that Mandarin listeners have similar RT trend to AusE listeners. That is, increasing duration results in lower RT for categorizing *horde*-based stimuli as /oː/ and higher RT for categorizing *hod*-based stimuli as /ɔ/. It shows that Mandarin learners are integrating acoustic cues and this process is important for L2 learners to approximate native-likeness.

5.2.2 Individual variation

Prior L2 vowel identification studies have shown large individual variability in using acoustic cues to perceive L2 vowels (Escudero et al., 2009; Mi et al., 2016). This study intended to examine the individual variability as well, but due to the time constraint for the project, this examination can only be conducted in later studies. However, based on descriptive analysis, duration has a variable effect on individual Mandarin learners' perception of /ɔ/-/oː/. Some Mandarin participants show exactly L1-like categorization, namely categorize the two vowels based on spectral properties only. Some participants are considerably influenced by duration, performing in the same way as they categorize /ɐ/-/ɐː/. It is hypothesized that the individual variability is correlated to length of study in Australia. Graphs that present individual categorization performance are included in the Appendix D to show the variability of perceptual responses in this study.

Chapter 6 Summary

The current study examined how vowel duration affects the perception of two AusE vowel contrasts /ɐ/-/ɐː/ and /ɔ/-/oː/ in both voiced and voiceless contexts by Mandarin learners of English. Sixty participants were recruited: 30 native Mandarin listeners who have been studied in Australia for at least six months and 30 native AusE listeners as control. Eight natural tokens containing the four vowels in /hVd/ and /hVt/ form were produced by a female native AusE speaker. Stimuli were created by varying the vowel duration of each natural token in 11 steps with endpoints 85ms and 335ms. With 4 vowels, 2 contexts, and 11 steps, this yielded 88 stimuli. Participants were asked to perform a two-alternative forced choice perceptual categorization task for each vowel contrast in each condition on those 88 stimuli. Categorization responses and RT data were collected. Descriptive analyses and statistical analyses using the general mixed-effects model were carried out.

The findings are: 1) both AusE listeners and Mandarin L2 learners rely on duration to categorize vowels as /ɐ/ and /ɐː/, but their category boundaries differ a little in location and differ significantly in steepness. Mandarin listeners have less steep boundaries, which along with their slower RT around the boundary, shows that they are less certain in their categorization than AusE listeners are. 2) Mandarin listeners are affected by duration and spectral quality when perceiving the /ɔ/ and /oː/ continua, whereas AusE listeners categorize those vowels largely based on the spectral feature. 3) both groups are influenced by voicing contexts when perceiving the /ɐ/ duration continuum. The category boundary is located at a longer duration step in the voiced than in the voiceless context.

The results imply that Mandarin learners of English can and will develop perceptual categories based on duration in the acquisition of English. This finding could support the hypothesis that L2 learners' reliance on duration cue in L2 sound perception is attributed to the development of duration categories during the L2 learning process.

References

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: The case of Japanese/r/and English/l/and/r/. *Journal of Phonetics, 32*(2), 233-250.

Baayen, R. H., & Milin, P. (2015). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12-28.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:doi:10.18637/jss.v067.i01

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

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *Language experience in second language speech learning: In honor of James Emil Flege, 1334*.

Blicher, D. L. (1988). Effects of syllable duration on the perception of Mandarin tones: A cross-language study. *The Journal of the Acoustical Society of America, 84*(S1), S157. doi:10.1121/1.2025902

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer [Computer Program]. Retrieved from http://www.praat.org/

Bohn, O.-S. (1995). Cross langauge speech productio in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 279-304). Baltimore: York Press.

Briere, E. J. (1966). An investigation of phonological interference. *Language, 42*(4), 768-796.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*(02), 201-251.

Cebrian, J. (2006). Experience and the use of non-native duration in L2 vowel categorization. *Journal of Phonetics, 34*(3), 372-387.

Cebrian, J. (2008). The effect of perceptual factors in the acquisition of an L2 vowel contrast. *Contrast in Phonology: Theory, Perception, Acquisition, 13*, 303-321.

Cebrian, J. (2008). Ocke-Schwen Bohn and Murray J. Munro, eds. 2007: Language Experience in Second Language Speech Learning: In honor of James Emil Flege.

Chen, H., Xu Rattanasone, N., Cox, F., & Demuth, K. (2015). *Effects of early dialectal exposure on adult perception of phonemic vowel length*. Paper presented at the Australian Linguistic Society Annual Conference 2015 (ALS2015), Western Sydney University, Sydney.

Cochrane, R. M. (1980). The acquisition of/r/and/l/by Japanese children and adults learning English as a second language. *Journal of Multilingual & Multicultural Development, 1*(4), 331-360.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavior Research Methods, Instruments, and Computers, 25*(2), 257-271.

Cox, F. (2006). The Acoustic Characteristics of /hVd/ Vowels in the Speech of some Australian Teenagers. *Australian Journal of Linguistics, 26*(2), 147-179. doi:10.1080/07268600600885494

Cox, F. (2012). *Australian English pronunciation and transcription*: Cambridge Cambridge University Press.

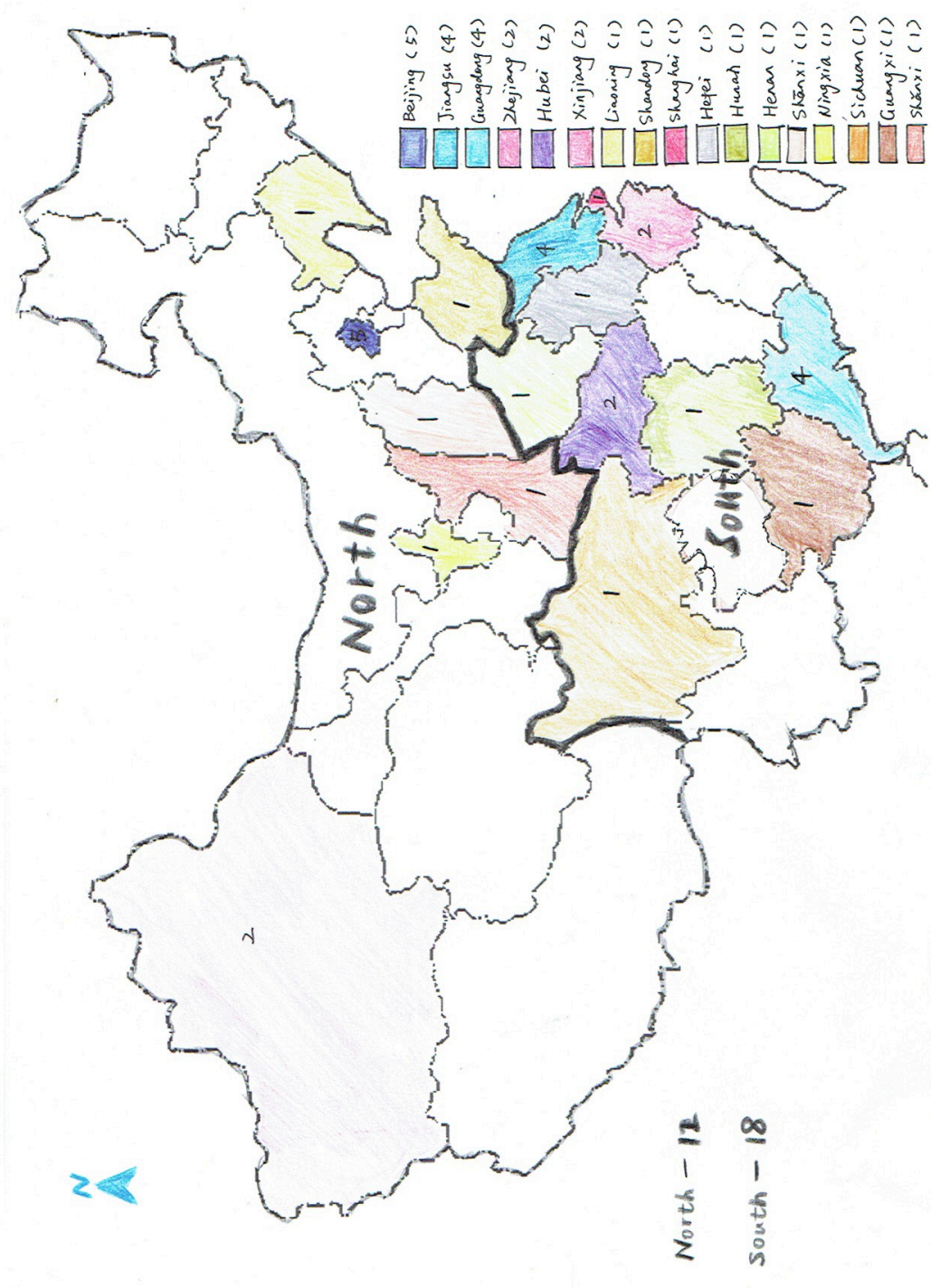
Cox, F., & Palethorpe, S. (2010). *Broadness variation in Australian English speaking females*. Paper presented at the Proceedings of the 12th Australasian International Conference on Speech Science and Technology.

Cox, F., Palethorpe, S., & Miles, K. (2015). *The role of contrast maintenance in the temporal structure of the rhyme*. Paper presented at the Proceedings of the 18th International Congress of Phonetic Sciences, Glasgow.
Cox, F. P. S. (2011). Timing differences in the VC rhyme of standard Australian English and Lenanese Australian English.
Escudero, P. (2009). Linguistic perception of "similar" L2 sounds. *Phonology in perception*, 152-190.
Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics, 37*(4), 452-465.
Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in second language acquisition, 26*(04), 551-585.
Escudero, P. R. (2005). Linguistic perception and second language acquisition : Explaining the attainment of optimal phonological categorization.
Fant, G. (1971). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations* (Vol. 2): Walter de Gruyter.
Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 233-277.
Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. *Phonetics and phonology in language comprehension and production: Differences and similarities, 6*, 319-355.
Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*(4), 437-470.
Flege, J. E., & Liu, S. (2001). The effect of experience on adult's acquistion of a second language. *Studies in second language acquisition, 23*(04), 527-552.
Gandour, J. (1977). On the Interaction between Tone and Vowel Length: Evidence from Thai Dialects. *Phonetica, 34*(1), 54-65. doi:10.1159/000259869
Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*(1), 110-125. doi:10.1037/0096-1523.6.1.110
Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia, 9*(3), 317-323. doi:10.1016/0028-3932(71)90027-3
Harrington, J., & Cassidy, S. (1999). *Techniques in speech acoustics* (Vol. 8): Springer Science & Business Media.
Harrington, J., Cox, F., & Evans, Z. (1997). An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Australian Journal of Linguistics, 17*(2), 155-184. doi:10.1080/07268609708599550
Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some Effects of Duration on Vowel Recognition. *Journal of the Acoustical Society of America, 108*(6), 3013-3022. doi:10.1121/1.1323463
Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones* (Vol. 6): Cambridge University Press.
Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *The Journal of the Acoustical Society of America, 59*(5), 1208.
Kondaurova, M. V., & Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *The Journal of the Acoustical Society of America, 124*(6), 3959-3971.

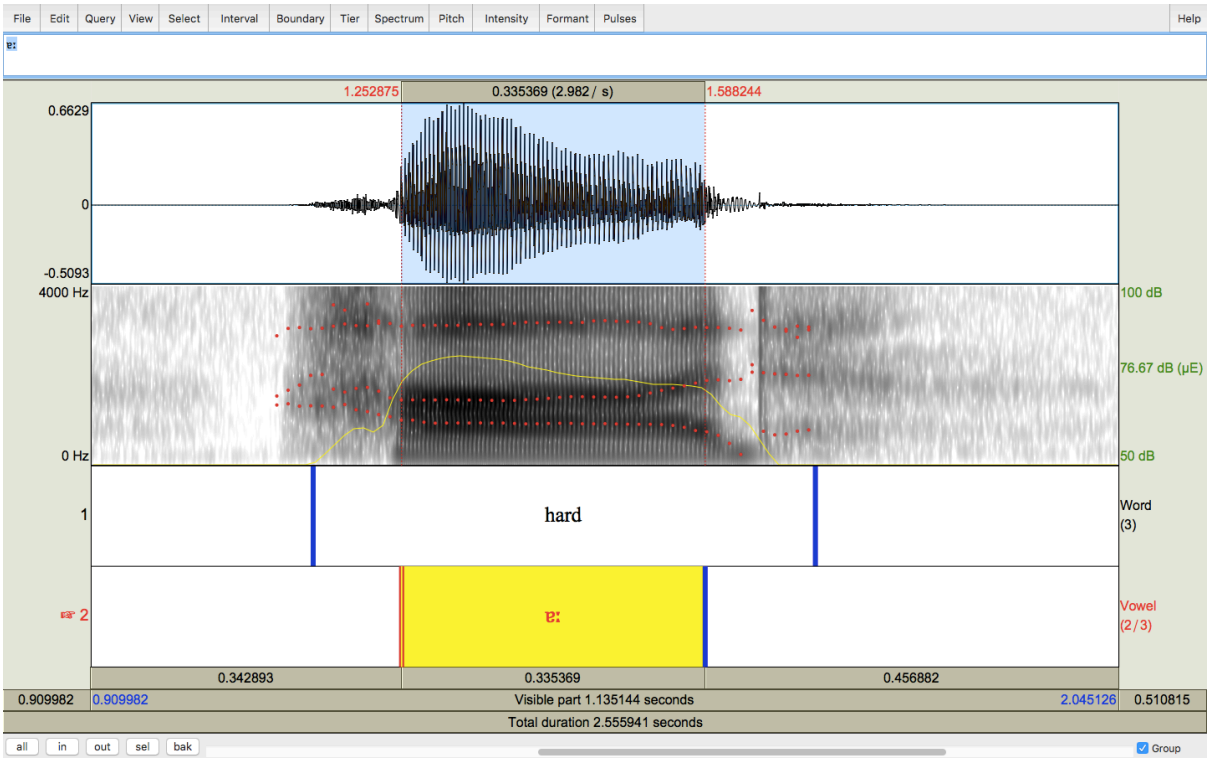
Kuhl, P. K., & Iverson, P. (1995). Chapter 4: Linguistic experience and the "Perceptual Magnet Effect". *Speech perception and linguistic experience: Issues in cross-language research*, 121-154.
Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*: Nelson Education.
Lee, W.-S., & Zee, E. (2003). Standard Chinese(Beijing). *Journal of the International Phonetic Association, 33*(1), 109-112.
Lehiste, I., & Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *The Journal of the Acoustical Society of America, 33*(4), 419-425.
Li, C., & Thompson, S. (1981). *Mandarin Chinese, a functional reference grammar* (Vol. 319). Los Angeles: Berkeley: University of California Press.
MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics, 2*(4), 369-390. doi:10.1017/S0142716400009796
Maddieson, I., & Disner, S. F. (1984). *Patterns of sounds*: Cambridge university press.
Massaro, D. W. (1987). Categorical partition: A fuzzy-logical model of categorization behavior. In S. Harnad (Ed.), *Categorical perception* (pp. 254-283). Cambridge, England: Cambridge University Press.
McAllister, R., Flege, J. E., & Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics, 30*(2), 229-258.
Mi, L., Tao, S., Wang, W., Dong, Q., Guan, J., & Liu, C. (2016). English vowel identification and vowel formant discrimination by native Mandarin Chinese- and native English-speaking listeners: The effect of vowel duration dependence. *Hearing Research, 333*, 58-65. doi:10.1016/j.heares.2015.12.024
Morrison, G. S. (2002). *Effects of L1 duration experience on Japanese and Spanish listeners' perception of English high front vowels* (Doctoral dissertation Doctoral dissertation), Simon Fraser University.
Morrison, G. S. (2007). Logistic regression modelling for first and second language perception data. In M. J. Solé, P. Prieto, & J. Mascaró (Eds.), *Segmental and prosodic issues in Romance phonology* (pp. 219 - 236). Amesterdam John Benjamins.
Morrison, G. S. (2009). L1-Spanish Speakers' Acquisition of the English/i/—/I/Contrast: Duration-based Perception is Not the Initial Developmental Stage. *Language and speech, 51*(4), 285-315.
Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15*(2), 285-290.
Pycha, A., & Dahan, D. (2016). Differences in coda voicing trigger changes in gestural timing: A test case from the American English diphthong/aɪ. *Journal of Phonetics, 56*, 15-37.
Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word‐final consonants in American English. *The Journal of the Acoustical Society of America, 51*(4B), 1296-1303.
Schneider, K., Dogil, G., & Möbius, B. (2011). *Reaction Time and Decision Difficulty in the Perception of Intonation.* Paper presented at the INTERSPEECH.
Selst, M. V., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The quarterly journal of experimental psychology, 47*(3), 631-650.
Stoicheff, M. L. (1981). Speaking Fundamental Frequency Characteristics of Nonsmoking Female Adults. *Journal of Speech Language and Hearing Research, 24*(3), 437. doi:10.1044/jshr.2403.437

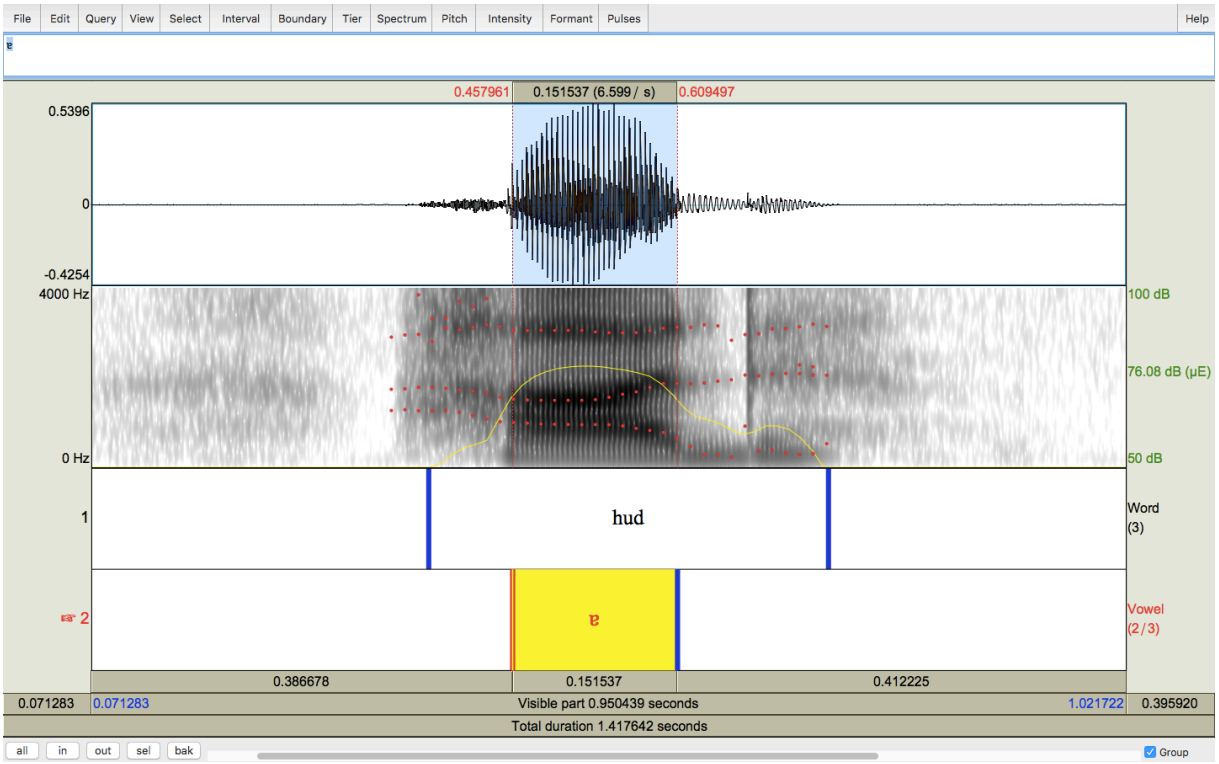
Strange, W. (1995). Cross-language study of speech perception: a historical review. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 3-45). Timonium, MD: York Press.
Strange, W. (2007). Cross-language phonetic similarity of vowels. *Language experience in second language speech learning: In honor of James Emil Flege, 17*.
Team, R. C. (2016). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.r-project.org/
Traunmüller, H., & Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. *Unpublished manuscript*.
Trubetzkoy, N. S. c. (1969). *Principles of phonology / [by] N. S. Trubetzkoy. Translated by Christiane A. M. Baltaxe*. Berkeley: Berkeley : University of California Press.
Tyler, M. D., Best, C. T., Faber, A., & Levitt, A. G. (2014). Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica, 71*(1), 4-21.
van Der Feest, S. V. H., & Swingley, D. (2011). Dutch and English listeners' interpretation of vowel duration. *The Journal of the Acoustical Society of America, 129*(3), EL57. doi:10.1121/1.3532050
van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: the L2LP model revised. *Frontiers in psychology, 6*.
Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *The Journal of the Acoustical Society of America, 106*(1), 458-468.
Woods, D. L., Wyma, J. M., Yund, E. W., Herron, T. J., & Reed, B. (2015). Factors influencing the latency of simple reaction time. *Frontiers in human neuroscience, 9*, 131.
Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the brain to weight speech cues differently: A study of Finnish second-language users of English. *Journal of Cognitive Neuroscience, 22*(6), 1319-1332.
Yu, A. C. (2010). Tonal effects on perceived vowel duration. *Laboratory phonology, 10*, 151-168.

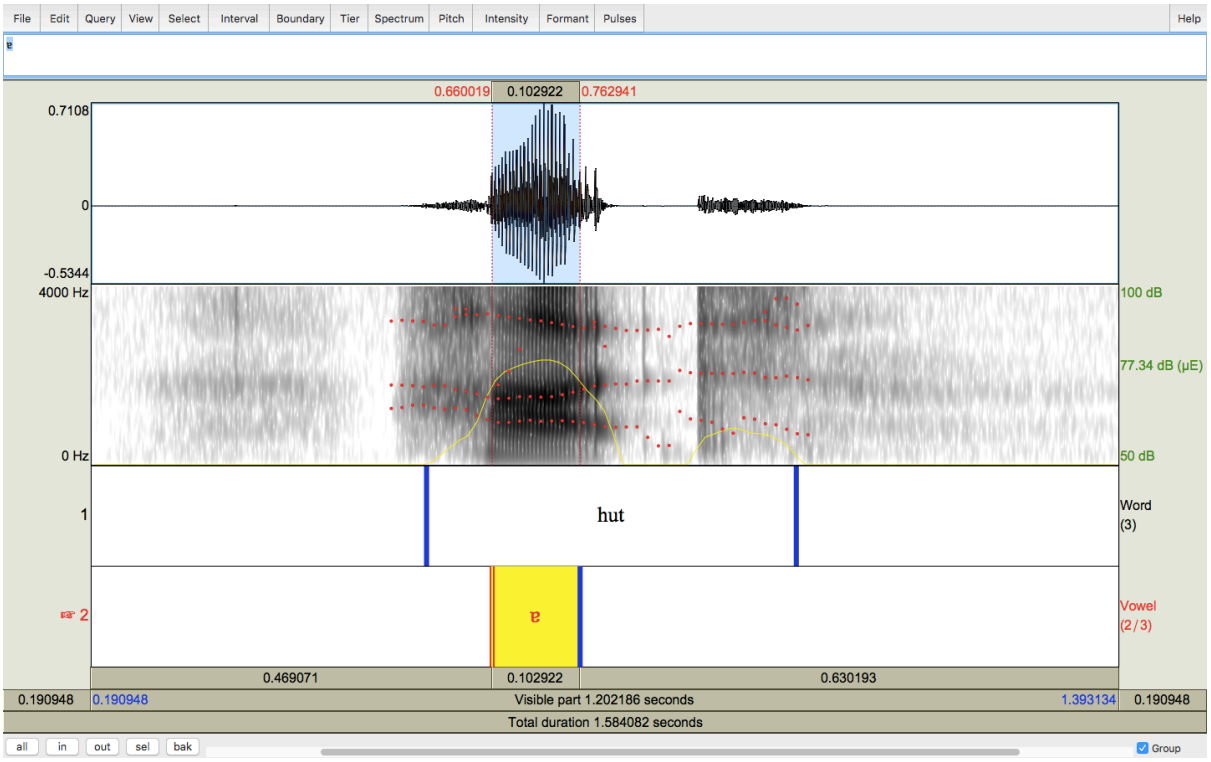
Appendix A The map of Mandarin participants' birth place

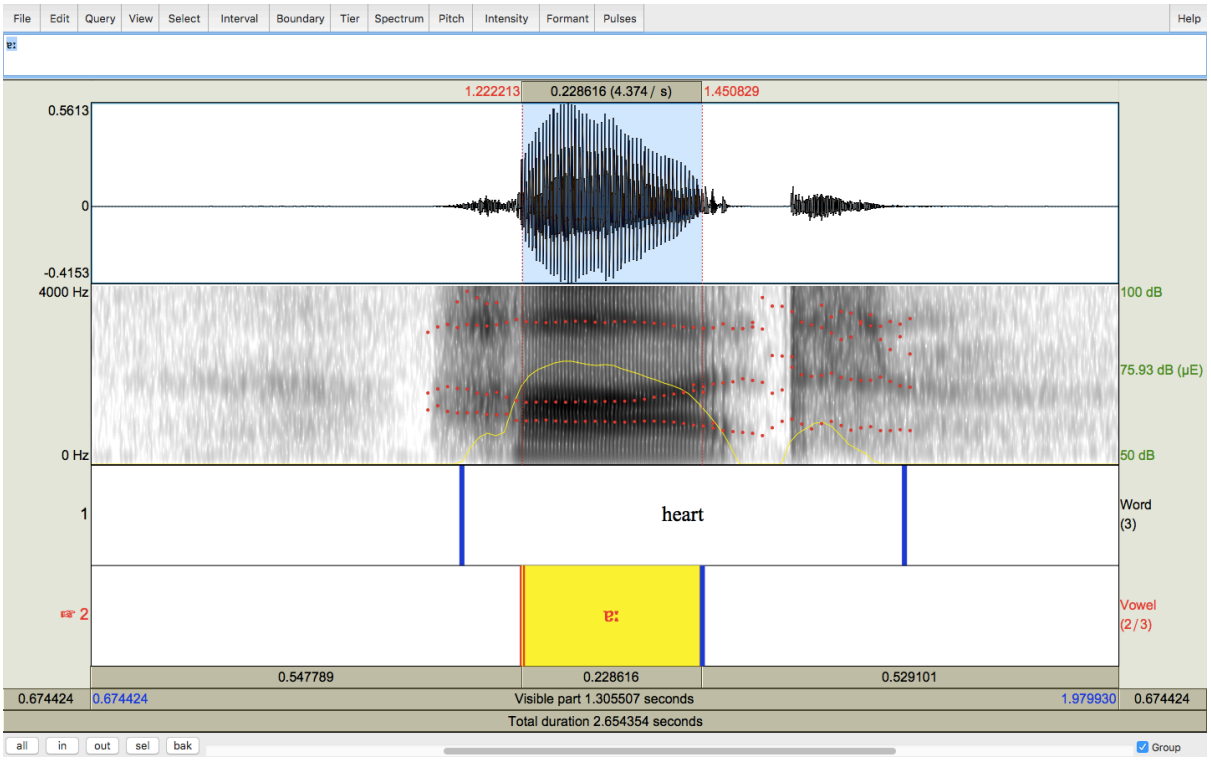


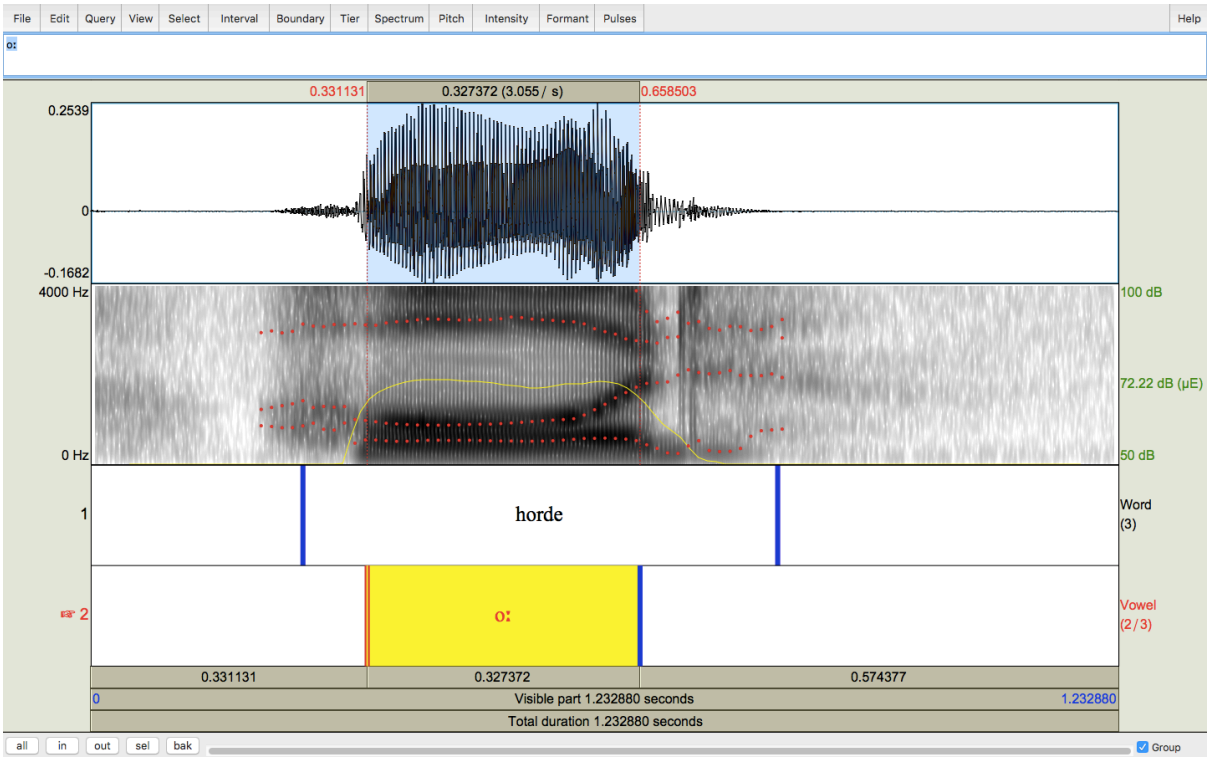
Appendix B Spectrograms of the eight natural tokens

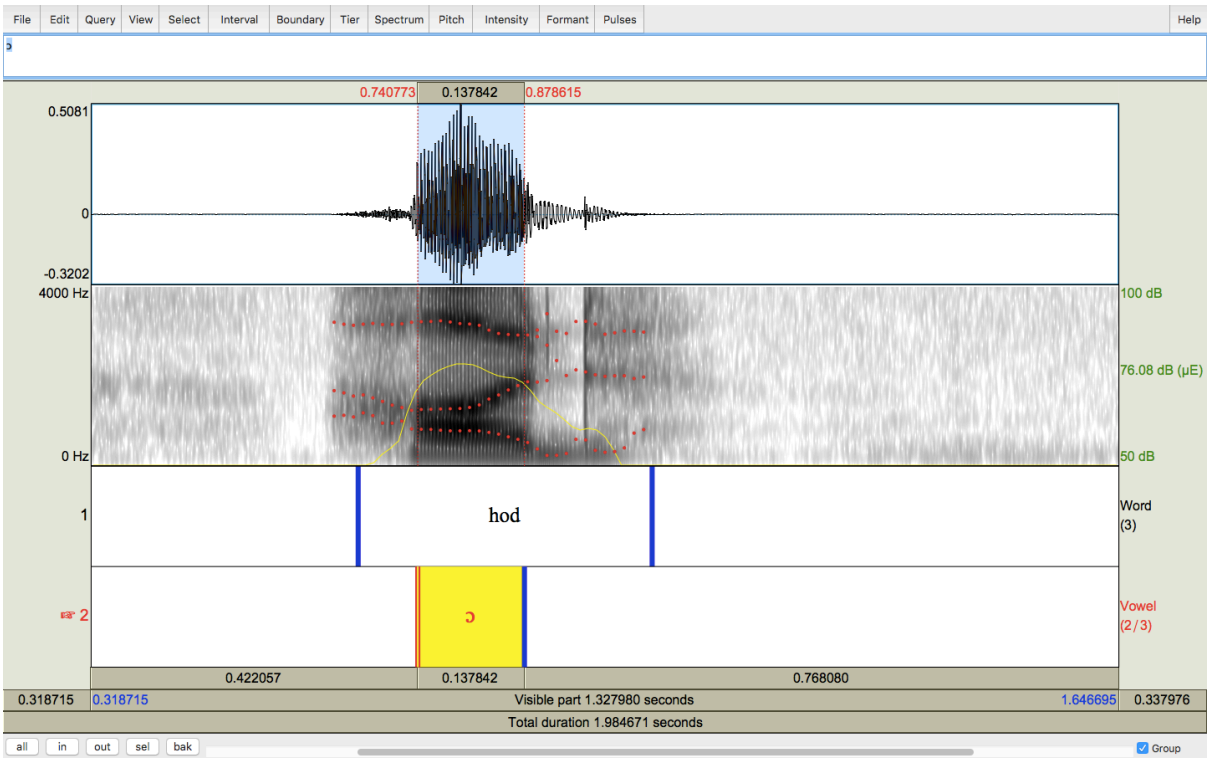


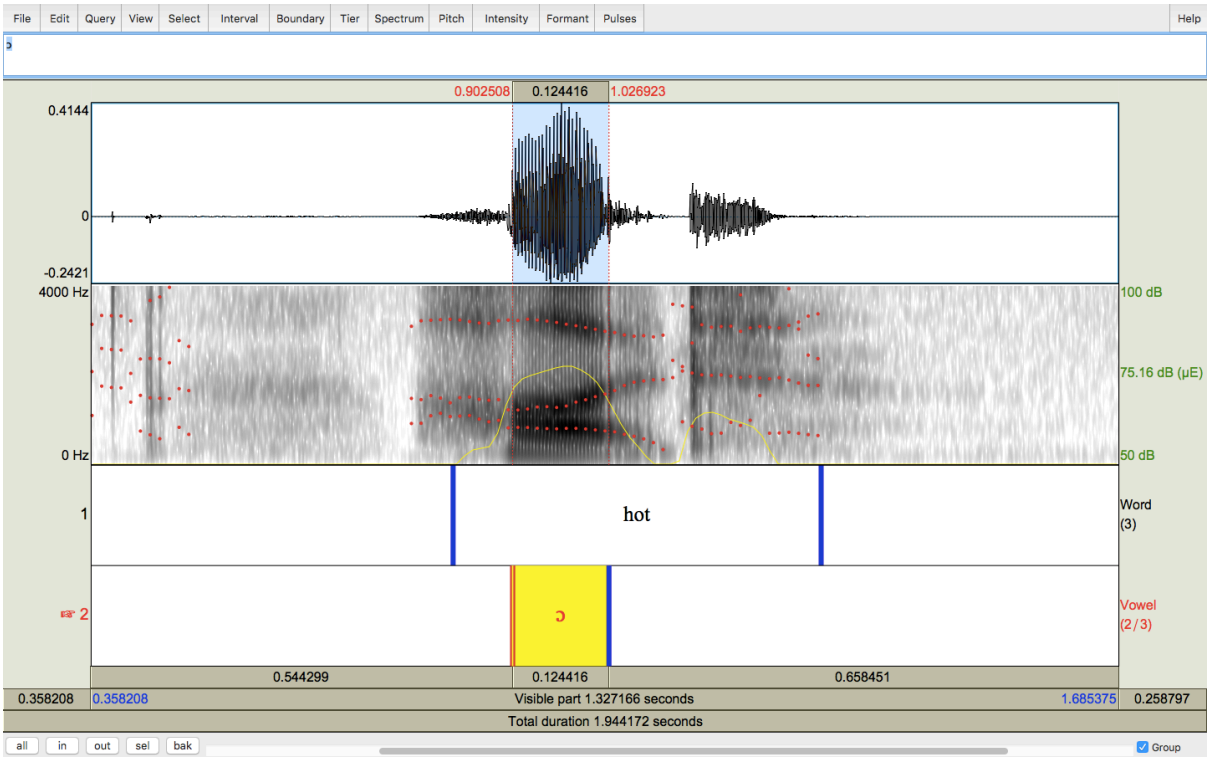


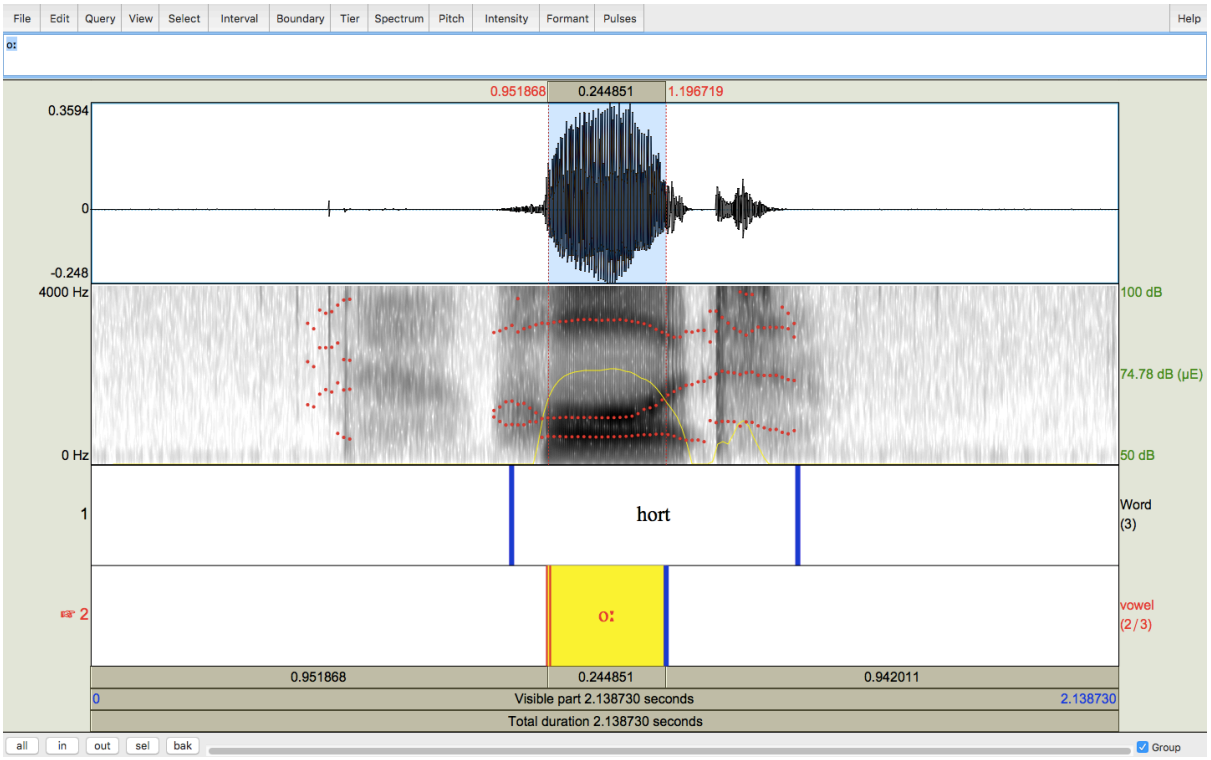

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
ɐ
0.660019 0.102922 0.762941
0.7108
0
-0.5344
4000 Hz
0 Hz
100 dB
77.34 dB (µE)
50 dB
1
hut
Word
(3)
2
ɐ
Vowel
(2/3)
0.469071 0.102922 0.630193
0.190948 0.190948 Visible part 1.202186 seconds 1.393134 0.190948
Total duration 1.584082 seconds
all in out sel bak
Group


File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
ɐː
1.222213 0.228616 (4.374 / s) 1.450829
0.5613
0
-0.4153
4000 Hz
0 Hz
100 dB
75.93 dB (µE)
50 dB
1
heart
Word
(3)
2
ɐː
Vowel
(2/3)
0.547789 0.228616 0.529101
0.674424 0.674424 Visible part 1.305507 seconds 1.979930 0.674424
Total duration 2.654354 seconds
all in out sel bak
Group

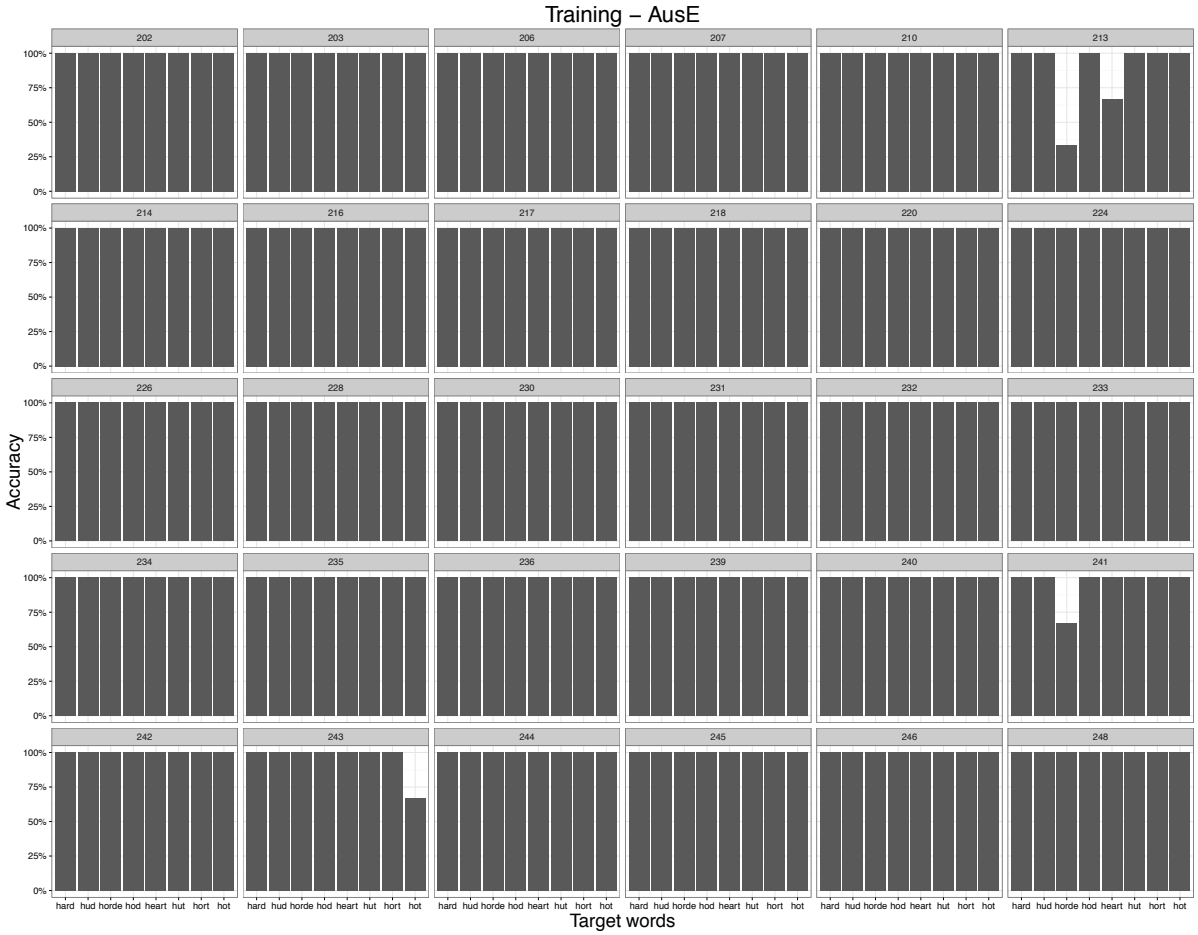

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
o:
0.331131
0.327372 (3.055 / s)
0.658503
0.2539
0
-0.1682
4000 Hz
0 Hz
100 dB
72.22 dB (µE)
50 dB
1
horde
Word
(3)
☞ 2
o:
Vowel
(2 / 3)
0.331131
0.327372
0.574377
0
Visible part 1.232880 seconds
1.232880
Total duration 1.232880 seconds
all in out sel bak
Group

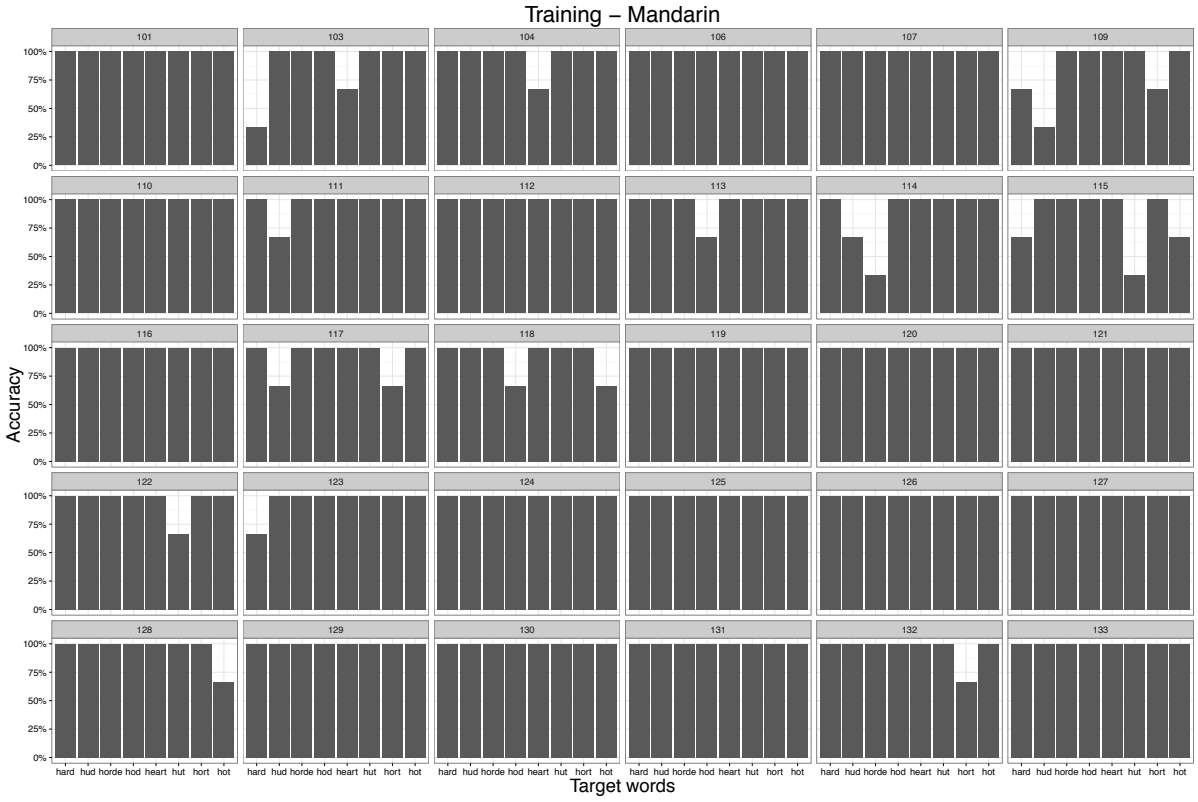

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
ɔ
0.740773
0.137842
0.878615
0.5081
0
-0.3202
4000 Hz
0 Hz
100 dB
76.08 dB (µE)
50 dB
1
hod
Word
(3)
☞ 2
ɔ
Vowel
(2 / 3)
0.422057
0.137842
0.768080
0.318715
0.318715
Visible part 1.327980 seconds
1.646695
0.337976
Total duration 1.984671 seconds
all in out sel bak
Group


File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
ɔ
0.902508 0.124416 1.026923
0.4144
0
-0.2421
4000 Hz
0 Hz
100 dB
75.16 dB (µE)
50 dB
1
hot
Word
(3)
☞ 2
ɔ
Vowel
(2 / 3)
0.544299 0.124416 0.658451
0.358208 0.358208 Visible part 1.327166 seconds 1.685375 0.258797
Total duration 1.944172 seconds
all in out sel bak
Group

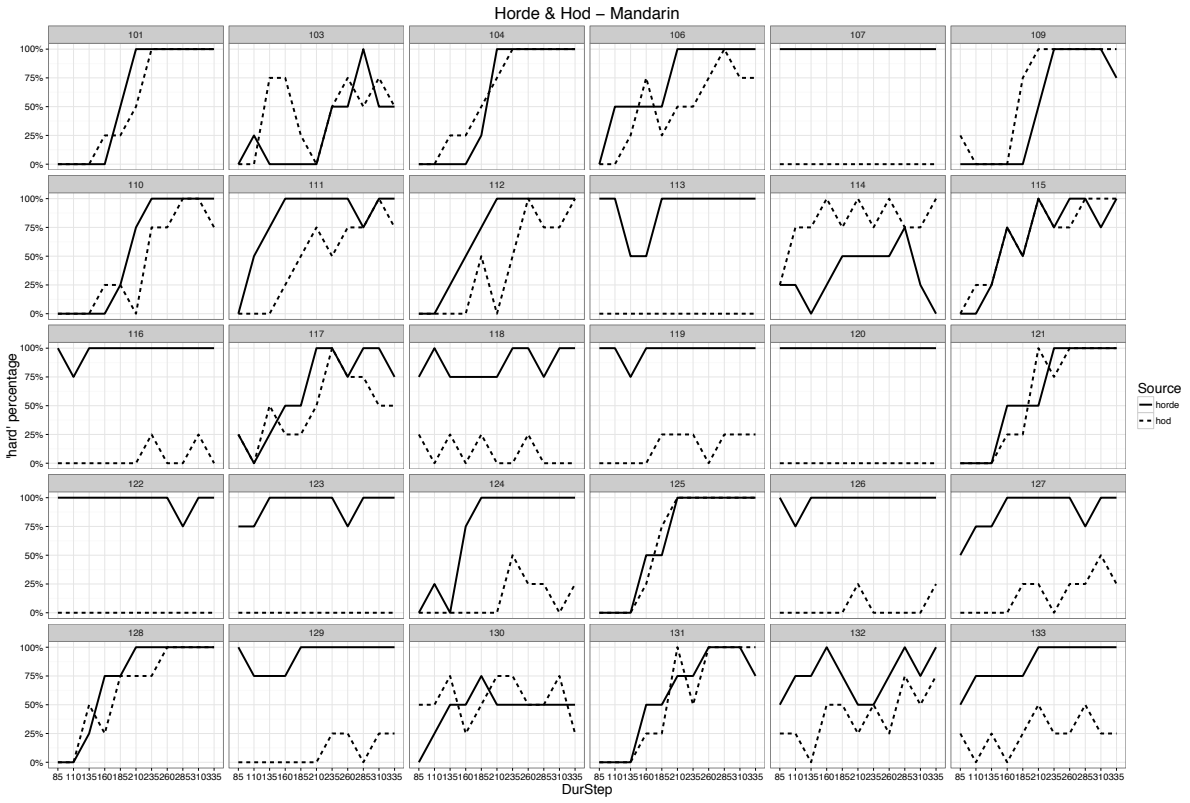

File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
oː
0.951868 0.244851 1.196719
0.3594
0
-0.248
4000 Hz
0 Hz
100 dB
74.78 dB (µE)
50 dB
1
hort
Word
(3)
☞ 2
oː
vowel
(2 / 3)
0.951868 0.244851 0.942011
0 Visible part 2.138730 seconds 2.138730
Total duration 2.138730 seconds
all in out sel bak
Group

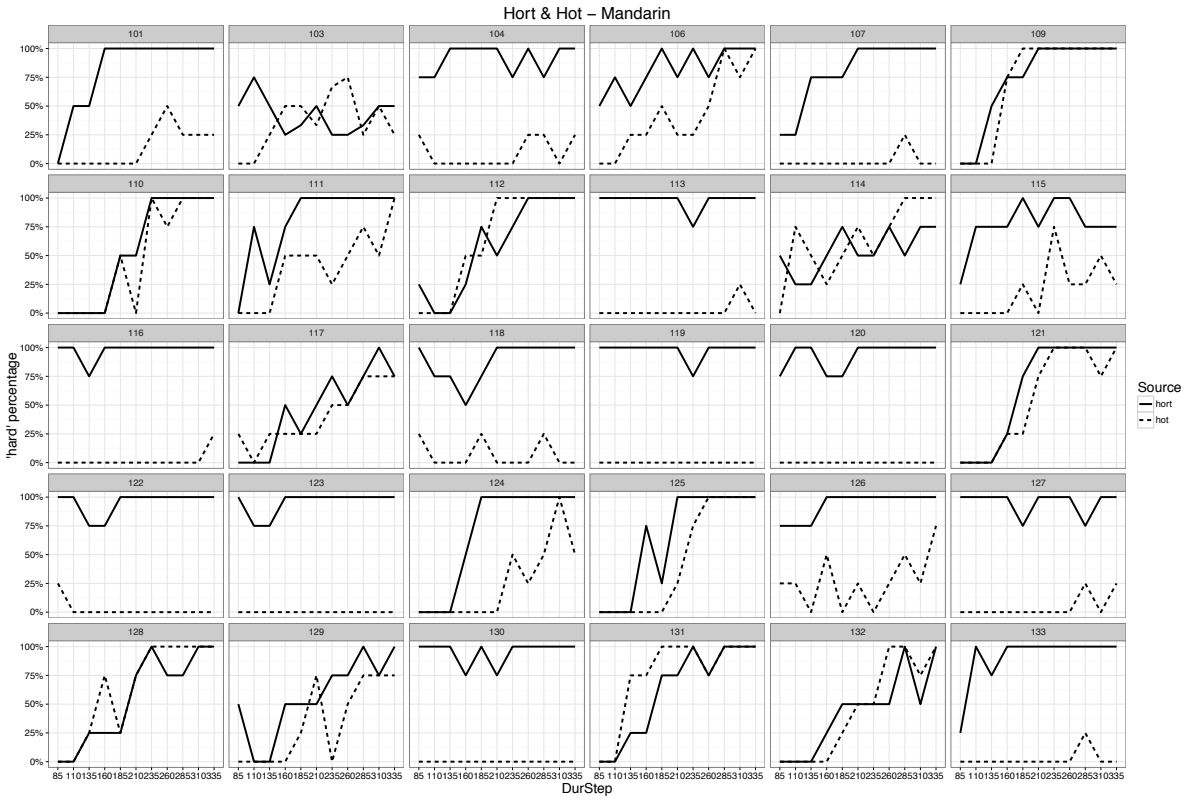
Appendix C Training data of both groups





Appendix D Mandarin group's individual performance on *hod-horde* and *hot-hort* block





Appendix E Ethics approval letter for the project

Fhs Ethics

RE: HS Ethics Application - Approved (5201600131)(Con/Met)

To: Associate Professor Felicity Cox Cc: Mrs Shuting Liu, Dr Titia Benders

Dear Associate Professor Cox,

Re: "Effects of Mandarin background on the perception and production of English speech sounds" (5201600131)

Thank you very much for your response. Your response has addressed the issues raised by the Faculty of Human Sciences Human Research Ethics Sub-Committee and approval has been granted, effective 31st March 2016. This email constitutes ethical approval only.

This research meets the requirements of the National Statement on Ethical Conduct in Human Research (2007). The National Statement is available at the following web site:

http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf.

The following personnel are authorised to conduct this research:

Associate Professor Felicity Cox
Dr Titia Benders
Mrs Shuting Liu

Please note the following standard requirements of approval:

1. The approval of this project is conditional upon your continuing compliance with the National Statement on Ethical Conduct in Human Research (2007).

2. Approval will be for a period of five (5) years subject to the provision of annual reports.

Progress Report 1 Due: 31st March 2017
Progress Report 2 Due: 31st March 2018
Progress Report 3 Due: 31st March 2019
Progress Report 4 Due: 31st March 2020
Final Report Due: 31st March 2021

NB. If you complete the work earlier than you had planned you must submit a Final Report as soon as the work is completed. If the project has been discontinued or not commenced for any reason, you are also required to submit a Final Report for the project.

Progress reports and Final Reports are available at the following website:

http://www.research.mq.edu.au/current_research_staff/human_research_ethics/application_resources

3. If the project has run for more than five (5) years you cannot renew approval for the project. You will need to complete and submit a Final Report and submit a new application for the project. (The five year limit on renewal of approvals allows the Sub-Committee to fully re-review research in an environment where legislation, guidelines and requirements are continually changing, for example, new child protection and privacy laws).

4. All amendments to the project must be reviewed and approved by the Sub-Committee before implementation. Please complete and submit a Request for Amendment Form available at the following website:

http://www.research.mq.edu.au/current_research_staff/human_research_ethics/managing_approved_research_projects

5. Please notify the Sub-Committee immediately in the event of any adverse effects on participants or of any unforeseen events that affect the continued ethical acceptability of the project.

6. At all times you are responsible for the ethical conduct of your research in accordance with the guidelines established by the University. This information is available at the following websites:

http://www.mq.edu.au/policy

http://www.research.mq.edu.au/for/researchers/how_to_obtain_ethics_approval/human_research_ethics/policy

If you will be applying for or have applied for internal or external funding for the above project it is your responsibility to provide the Macquarie University's Research Grants Management Assistant with a copy of this email as soon as possible. Internal and External funding agencies will not be informed that you have approval for your project and funds will not be released until the Research Grants Management Assistant has received a copy of this email.

If you need to provide a hard copy letter of approval to an external organisation as evidence that you have approval, please do not hesitate to contact the Ethics Secretariat at the address below.

Please retain a copy of this email as this is your official notification of ethics approval.

Yours sincerely,

Dr Anthony Miller
Chair
Faculty of Human Sciences
Human Research Ethics Sub-Committee

Faculty of Human Sciences - Ethics
Research Office
Level 3, Research HUB, Building C5C
Macquarie University
NSW 2109

Ph: +61 2 9850 4197
Email: fhs.ethics@mq.edu.au
http://www.research.mq.edu.au/