May 2016

Covert Contrast: the Acquisition of Mandarin Tone 2 and Tone 3 in L2 Production and Perception

Liya Mar
University of Wisconsin-Milwaukee

COVERT CONTRAST: THE ACQUISITION OF MANDARIN TONE 2 AND TONE 3 IN L2 PRODUCTION AND PERCEPTION

by

Li-Ya Mar

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Linguistics

at

The University of Wisconsin-Milwaukee

May 2016

ABSTRACT

COVERT CONTRAST: THE ACQUISITION OF MANDARIN TONE 2 AND TONE 3 IN L2 PRODUCTION AND PERCEPTION

by

Li-Ya Mar

The University of Wisconsin-Milwaukee, 2016
Under the Supervision of Professor Fred Eckman

This dissertation investigates the occurrence of an intermediate stage, termed a covert contrast, in the acquisition of Mandarin Tone 2 (T2) and Tone 3 (T3) by adult speakers of American English. A covert contrast is a statistically reliable distinction produced by language learners that is not perceived by native speakers of the target language (TL). In second language (L2) acquisition, whether a learner is judged as having acquired a TL phonemic contrast has largely depended on whether the contrast was perceived and transcribed by native speakers of the TL. However, categorical perception has shown that native listeners cannot perceive a distinction between two sounds that fall within the same perceptual boundaries on the continuum of the relevant acoustic cues. In other words, it is possible that native speakers of the TL do not perceive a phonemic distinction that is produced by L2 learners when that distinction occurs within a phonemic boundary of TL

The data for the study were gathered through two elicitations of tone production, a longitudinal analysis, and two perception tasks. There were three key findings. First, both elicitations showed that most of the L2 participants produced a covert contrast between T2 and T3 on at least one of the three acoustic measures used in the study.

Second, the longitudinal analysis reveals that some L2 participants progressed from making a covert contrast to a later stage of implementing an overt one, thereby supporting the claim that making a covert contrast is an intermediate stage in the process of acquiring a L2 phonemic contrast. Third, results of the perceptual tasks showed no reliable difference in identifying and discriminating Mandarin T2 and T3 on the part of the L2 learners who produced a covert contrast and those who produced an overt contrast, indicating that there was no reliable difference in the two groups' ability to perceive the target tones.

In all, the occurrence of a covert contrast in the process of acquiring Mandarin T2 and T3 suggests that L2 acquisition of a tonal contrast is a gradient process, one in which an intermediate step occurs before a L2 learner reaches the final stage of implementing an overt contrast that is perceived as target-like by the native speakers of the TL.

To my family,

and all people who share love and compassion

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SLA	Second Language Acquisition
L2	Second Language
TL	Target Language
L1	First Language
IL	Interlanguage
T1	Tone 1 (in Mandarin)
T2	Tone 2 (in Mandarin)
T3	Tone 3 (in Mandarin)
T4	Tone 4 (in Mandarin)
VOT	Voice Onset Time
PAM	Perceptual Assimilation Model
SLM	Speech Learning Model
F0	Fundamental Frequency
OPI	Oral Proficiency Interview
ACTFL	American Council of the Teaching of Foreign Languages
CV	Consonant-Vowel
Hz	Hertz
NT2	Non-T2
NT3	Non-T3
RT	Reaction Time
ANOVA	Analysis of Variance
NS	Native Speakers

ACKNOWLEDGMENTS

There are many people I would like to thank. Without them, I would not have been able to complete this dissertation and my doctoral study. First and foremost, I am sincerely grateful to Professor Fred Eckman, who is more than an advisor, but a mentor and a wise friend who has supported me and always wanted the best for me throughout my time in Milwaukee. Professor Eckman has always kept his door open for me whenever I needed guidance. He has spent endless hours reading my drafts and patiently writing me recommendation letters, and what I'm most grateful for is that he has never shown "I'm very busy" on his face in any instances when I sought guidance or support from him. I am inspired not only by his passion for research in SLA and Linguistics, but also by his dedication to helping his students and advisees succeed. With such dedication and positive approach, Professor Eckman has become my role model and will always be a great influence in my teaching career. Thank you, Fred.

I would like to express my gratitude to my committee members, Professor Yea-fen Chen, Anne Pycha, and Jae Yung Song. Professor Chen was the first person who believed in me and offered me my first teaching job as a master's student, which eventually brought me to UWM, a place where I've grown as a linguist and a teacher. I truly appreciate that she has continued to provide me with guidance in teaching Mandarin and in the pedagogical aspects of my research career. Professor Pycha has been very supportive throughout my study here. She always gave detailed and insightful feedback on the statistical and acoustic aspects of my study, and challenged me with questions that I never thought about before, and which, in turn, have tremendously helped shape my critical thinking ability. Her active involvement in the Linguistics

Department has brought the entire program closer together and helped form a very supportive linguistic community. Professor Song has always been very encouraging and helpful during the process of completing my dissertation and my doctoral studies. She always welcomed me to her office with a genuine smile, and helped answer my questions with extreme patience. She walked me through every detail of the statistical methods that I did not understand and always provided me with all the resources that she has in order to solve the problems I encountered. I am deeply thankful for the effort and time my committee members have put into helping me complete this piece of scholarship. Without their support and their challenges to my work, I wouldn't have been able to complete this dissertation.

Next, I want to thank Professor Hanyong Park, who was the first person to introduce acoustic and auditory phonetics to me. The opportunity to work with him on his acoustic projects in 2012 and 2013 has greatly helped me with my understanding and application of acoustic analysis in SLA research. Professor Greg Iverson, my advisor for my master's study is another person I want to thank. He was the first person who encouraged me to pursue a doctoral degree. I also want to express my appreciation to all the faculty members and staff in the Linguistics Program and the Chinese Language Program, for their encouragement and support throughout the six years. I very much enjoyed working with the kind and genuine people in these two programs, which has made my life and academic career in Milwaukee more fun and memorable.

I also want to thank my undergraduate advisors in Taiwan, Professor Hui-chun Hsu, Mei-chun Liu, Chen-sheng Liu. They were the very first people who introduced me

to Linguistics and helped me find my path to eventually pursue Linguistics for my academic career.

In addition, I am thankful for my friends, my fellow graduate students, and teaching assistants, especially Vivian Chiang, Laurel Schenkoske, Siriporn Lerdpaisalwong, Maia Stack, Meghan Murphy, Yingfei Chen, Yu-chun Lin, Xiaorong Wang, Andrew Olson, Zafer Lababidi, Sooyeon Lee, Ho Eun Park and Kelsie Miller for encouraging me, giving me advice, and sometimes listening to me complain when things got frustrating. They have been my support group at Milwaukee and I am truly happy to have met them here. I also want to thank my students of Chinese, some of whom participated in my study and became my friends. They have helped me find the topic of the dissertation and inspired me to be a better educator everyday.

Lastly, yet most importantly, I'd like to thank my family. I am especially and deeply grateful to my mother, Hui-lin Wang, my husband, Jeffrey Lu, and my younger brother, Jia-cheng Ma. Their endless encouragement, cheerful spirit, and unconditional love have carried me through the ups and downs, laughs and tears of this fun adventure. I am also writing this in the loving memory of my father, Yuan-heng Ma, who passed away three years ago.

Chapter 1

Introduction

In the body of research on second language acquisition (SLA), whether a second language (L2) learner has acquired the production of an L2 phonemic contrast has largely depended on whether the contrast has been perceived by native speakers of the target language (Flege, 1987; Eckman & Iverson, 1994). That is, if native speakers do not perceive the contrast as produced by the L2 learners, it is assumed that the L2 learners have not successfully produced and acquired the contrast. However, it has been demonstrated that native listeners of a language perceive phonemic differences categorically, such that within a given category there can be variation in the relevant acoustic cues as long as the phonemic boundary is not crossed (Liberman et al, 1957; Goldinger et al, 1996). In other words, it is possible that L2 learners are producing a reliable, systematic distinction between two target language (TL) phonemes where native speakers do not perceive these sounds as belonging to distinct categories, because the acoustic differences produced by the L2 learners fall within a single phonemic category. In this case, the L2 learners are not perceived as making the target-like contrast between the two sounds in question.

Some studies in first language (L1) acquisition have shown that in the process of acquiring their native language, children produce systematic contrasts that are not perceived by adult listeners (Macken & Barton, 1980; Scobbie, Gibbon, & William, 1996; Li, Edwards, & Beckman, 2009). Some of these distinctions produced by children have been shown to be statistically reliable, and illustrate the development of the acquisition of phonemic categories in L1. This phenomenon in which a statistically significant

difference is produced by the learners, but not perceived by native-speaking listeners has been referred to in the field of L1 acquisition as a “covert contrast”,

Recently, the idea of covert contrasts has been investigated in SLA to see whether they occur in the process of acquiring target language (TL) phonemes. Eckman, Iverson and Song (2014) conducted a study on the acquisition of the English /s/-/z/ contrast by native speakers of Spanish. The study showed that native speakers of English in some cases did not hear the contrast as produced by several of their L2 participants, who were nevertheless producing a statistically reliably distinction between the /s/ and /z/ targets. This finding indicates that while the distinction that some of the L2 learners produced was not implemented in a target-like fashion, and thus not overtly perceived by the native speakers, these participants were implementing a systematic distinction between /s/ and /z/ in their own interlanguage (IL). What the study also suggests is that since a systematic distinction has been implemented in their IL, the participants could eventually make further progress to where they implement a target-like contrast. In other words, native speakers’ perception of L2 production is not always reliable in determining whether a L2 learner is able to produce a TL contrast. The present study aims to investigate the L2

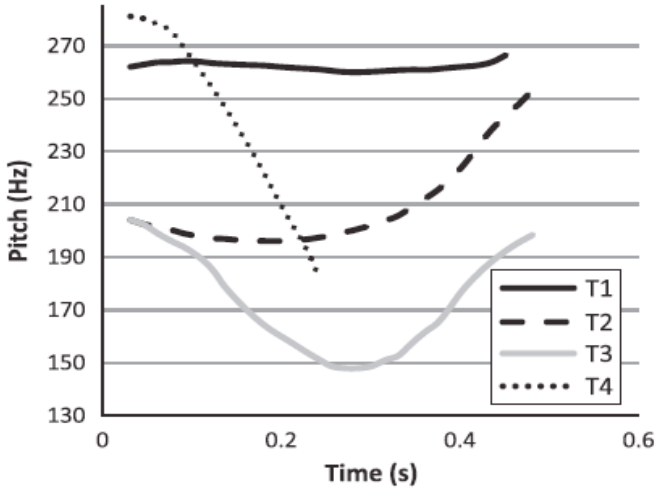


Figure 1.1: The pitch contour of Mandarin four tones produced by a female native speaker in the study by Hao (2012)

acquisition of the contrast between two of the tones in Mandarin. Mandarin is a tonal language which has four lexical tones (T1, T2, T3, and T4). Figure 1.1 shows the pitch contours of the four tones produced by a female Mandarin speaker in the study conducted by Hao (2012) . The purpose of this study is to investigate whether L2 learners of Mandarin produce a covert contrast between T2 and T3. Overall, tonal contrasts have been observed as the most difficult aspect in the L2 learning of Mandarin. Many studies have reported that among the four lexical tones in Mandarin, T2 and T3 are found to be the most difficult to discriminate and to produce, for both native speakers and L2 learners (Shih et al, 2010; Zhang, 2011; Hao, 2012; Shen et al, 2013). This difficulty is presumably due to the similarity of the pitch contour of T2 and T3. As shown in Figure 1, T2 is an overall rising tone in which the pitch falls slightly at the beginning, and then rises from about 205 Hz to 250 Hz at the contour offset. T3 is a "dip-rise" tone which falls significantly at the beginning from 205 Hz to 150 Hz before rising to 205 Hz at the end. Since the pitch for both T2 and T3 rises at the end of the contour, it is easy to confuse the two, especially in fast speech in which parts of the contour are sometimes omitted or shortened by native speakers.

Although a number of studies, as cited above, have investigated the L2 acquisition of T2 and T3, the current study is intended to extend this reseach by investigating whether L2 learners produce a covert contrast in the acquisition of these tones. In fact, there have been only a few studies on covert contrast in SLA, and the study by Eckman et al (2014) was one of the first to have addressed the L2 production of such a contrast. The current study aims to continue the investigation of the

occurrence of covert contrast in L2 acquisition by adding data from the learning of Mandarin tones. Also, if, as proposed in L1 acquisition (Macken & Barton, 1980), covert contrast is considered a stage in the acquisition process, an interesting question would be whether L2 learners would eventually move from making the contrast covertly to implementing the contrast overtly. Attempting to answer this question requires the collection of longitudinal data. No published study is known to have conducted a longitudinal analysis to investigate the development from a covert contrast to an overt contrast in L2 production. This present study intends to address this question by gathering longitudinal data on the L2 acquisition of Mandarin tones. Moreover, many studies have discussed the relationship between perception and production in L2 acquisition, but no studies have looked at the perceptual ability of L2 learners who make a TL contrast covertly. It will be an empirically interesting question to investigate whether the perceptual ability of L2 learners who implement a covert contrast differ from the perception of L2 learners who make an overt contrast or no contrast. Thus, this study also aims to contribute to the body of research on the relationship between L2 perception and production by analyzing both production and perception data from L2 learners who evidence a covert contrast, specifically in the acquitision of Mandarin T2 and T3.

To recapitulate briefly, the purposes of the study are the following. First, it aims to report the findings of an investigation of L2 production of Mandarin T2 and T3 and to determine whether English-speaking L2 learners of Mandarin produce a covert contrast in the process of acquiring these tones. Secondly, it will report the results of a

longitudinal study to document whether a L2 learner may progress from making a covert contrast to a later stage of implementing an overt one, thereby testing the claim that making a covert contrast is an intermediate stage in the process of acquiring a L2 phonemic contrast. Thirdly, this study reports the findings of two perceptual tasks on Mandarin T2 and T3, with the intent of contributing to the discussion on the relationship between L2 perception and production. More specifically, this study will attempt to determine whether there is any difference in perceiving Mandarin T2 and T3 between the L2 learners who produce a covert contrast and those who produce an overt contrast.

The findings of the study are intended to help answer the ultimate big-picture question of whether L2 phoneme acquisition is a gradient process, in which there may be an intermediate stage in the IL where L2 learners implement a systemic yet not TL-like contrast before reaching the TL-like stage. If covert contrast is found to be an intermediate stage in L2 acquisition, then language acquisition would appear to be a gradient process, one in which an intermediate step occurs between a stage of no contrast and a stage in which L2 learners produce a TL-like contrast. In addition, the criterion for whether a L2 learner has acquired a phonological contrast should not depend solely on whether the contrast is perceived as target-like by native speakers, but instead should be determined at least in part by what the learner is producing.

The dissertation is structured as follows. Chapter 2 identifies the question that the current study tries to answer, and provides the context of the current study by reviewing what has been found and discussed in other studies on covert contrast, on the distinction between T2 and T3, and on L2 production and perception. The big-

picture question and the current hypothesis in the attempt to answer it, along with the specific hypotheses for the present study are laid out at the end of Chapter 2. Chapter 3 presents the production experiment with detailed descriptions on its methodology and results, including the discussion of the longitudinal data. Chapters 4 and 5 introduce each of the two perception experiments. Chapter 4 concerns the methods emplyed to gather and analyze data of the identification task of the two Mandarin tones whereas Chapter 5 lays out the details of the discrimination task. Finally, Chapter 6 provides a complete discussion by linking the results of all the experiments and addressing the hypothesis posted and concludes this dissertation.

Chapter 2

Background

The objective of this chapter is to introduce relevant background information about L2 production and perception of Mandarin T2 and T3, and to connect to what has been done in the field of L2 acquisition with respect to how L2 production is evaluated. This study aims to document the occurrence of a covert contrast in L2 acquisition, something that has been attested only sparsely and recently. This section will be presented in the following order: first, Section 2.1 reviews previous findings on covert contrast in L1 and L2 acquisition. Second, the acoustic measurements that differentiate Mandarin T2 and T3 in production are laid out. Section 2.3 describes the acoustic cues that are employed in perceiving Mandarin tones. Sections 2.2 and 2.3 are crucial in discussing and interpreting the L2 tonal data gathered in the study. Before stating the hypotheses that are being tested, I present a discussion in 2.4 on the long-debated relationship between L2 perception and production. One of the goals of this study is to contribute to this debate. Finally, the hypotheses that I aim to test are stated in Section 2.5.

2.1 Covert Contrast

The term “covert contrast” is used to refer to the production of a statistically reliable acoustic distinction produced by a language learner that is not perceived by native speakers. The idea of covert contrast was first documented in the field of first language acquisition (FLA) by Macken and Barton (1980) who measured the VOT of

word-initial stops as produced by young children acquiring English. VOT is one of the most frequently used acoustic cues to distinguish the voicing and aspiration of stops in the world's languages (Lisker & Abramson, 1964). The children in Macken and Barton's study were able to produce statistically reliable VOT distinctions in the target stops, but those distinctions all fell within the VOT category of what adult speakers produced as voiced stops.

More specifically, Macken and Barton identified three stages shown in young children acquiring a VOT distinction in English word-initial stops. In the first stage, the children produced no systematic VOT distinction at all. In the second stage, children produced a statistically reliable VOT difference between the voiced and voiceless targets, however these VOT values all fell within the phonetic category that native speakers of English perceived as a voiced stop. In the third stage, the children produced VOT differences between the voiced and voiceless targets that were in the same range as those produced by adult native speakers.

The study by Li et al (2009) investigated the phonological development of toddlers acquiring the contrast between voiceless alveolar and post-alveolar fricatives in English and in Japanese. Covert contrasts were also observed in the frequency of the vowel following the fricatives for both groups of toddlers despite the fact that the acquisition sequence for the two language groups was the opposite.

In the study of L2 acquisition of English /s/-/z/ contrast as produced by native speakers of Spanish, Eckman et al (2014) found that some of their participants produced a statistically reliable difference in the percent of the overlap between the

voicing and frication of the /s/ and /z/ targets. The overlap between the voicing and frication was significantly longer for the /z/-targets than for the /s/-targets, however this distinction was not perceived by the native speakers of English. Although Eckman et al did not obtain results that would fully support the occurrences of the three stages identified by Macken and Barton, they suggested that if their results can be interpreted similarly to the ones found in research on L1 acquisition and on disordered speech, the three stages would be expected to occur in the process of acquiring a L2 contrast.

The acquisition stages suggested by Macken and Barton, and by Eckman et al are the direction that the present study hopes to pursue. No study has yet reported investigating whether there is a stage of covert contrast in the development of the L2 acquisition of lexical tones in Mandarin. Neither has any study provided longitudinal analysis in supporting the occurrence of such an intermediate stage of covert contrast and how L2 learners proceed in the acquisition process. The goal of the present study is to examine whether the stage of covert contrast occurs in the T2-T3 contrast from L2 learners' tonal production. If a covert contrast is attested in the L2 production of the T2-T3 contrast, this would suggest that an intermediate stage of convert contrast is not limited only to segments, but can involve suprasegmental levels as well, as in the tonal level of Mandarin. The hypotheses that the present study aims to test are stated in detail after all background information is introduced.

2.2 The T2-T3 Contrast in Production

From the literature, covert contrast can primarily be attested on the continuum of an acoustic measure. The study by Moore and Jongman (1997) indicates that the acoustic measures used to distinguish all four Mandarin tones include primarily both the F0 contour and F0 height. T2 and T3 in Mandarin are both contour tones in which the pitch changes to complete the tones. Although T2 and T3 are confusable because of their similar rising direction in the later half of the contour, according to Shen (1990) and to Moore and Jongman (1997), the two tones are distinguishable in the F0 difference between the onset of the contour and the turning point. As F0 in both T2 and T3 falls in the beginning and then rises later in the contour, the turning point refers to the point where F0 stops falling and starts rising, usually the point with the lowest F0. The measures of F0 difference and duration to the turning point are illustrated in Figure 2.1, taken from Moore and Jongman (1997). As shown in Figure 1.1 (repeated here for

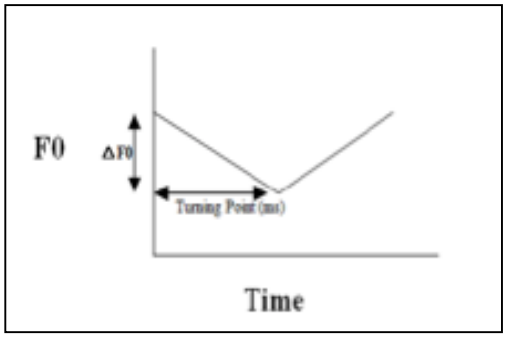


Figure 2.1: The acoustic measures that distinguish T2 and T3: The F0 difference between the contour onset and the turning point and the time it takes to the turning point

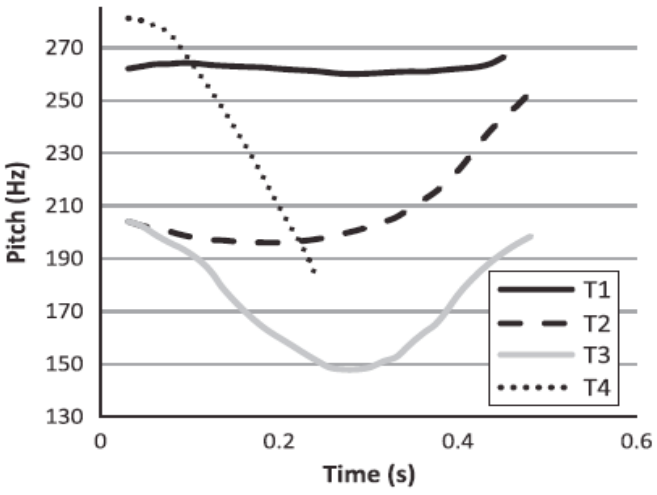


Figure. 1.1: The pitch contour of Mandarin four tones produced by a female native speaker in the study by Hao (2012)

convenience), the F0 difference between the contour onset and the turning point is greater in T3 than in T2. T2 has a smaller drop in F0 before the turning point than does T3. Also, previous studies (Shen, 1990; Zhang, 2011) show that the duration (millisecond) to the turning point in T2 and T3 is reliably different. It takes T3 longer to reach the turning point than it does T2.

In a previous pilot study, the target tones were elicited in naturally-produced sentences where the F0 difference was the only acoustic measure used. This is because the durational cue cannot be clearly measured in context where the speech rate plays a role. Also, the measures of the F0 cues were affected by the neighboring tones, because coarticulation occurs in tones in context, as shown by several previous studies (Shen, 1990; Xu, 1994; Jongman et al, 2006). In the current study, target tones have been elicited in isolation without any proceeding or following tones, which enabled including the durational cue of time to the turning point in the analysis and exclude any tone sandhi triggered by the contexts.

In addition to the two acoustic cues discussed above, offset pitch height has been shown to be another measure that systematically differentiates T2 from T3. Shen et al (2013) used eye-movement techniques to investigate the acoustical cues for identifying synthetic T2 and T3 by native speakers of Mandarin; they found that the judgment on T2 and T3 were determined on the basis of the offset pitch height. Targets with a high F0 at the end of the contour were systematically identified as T2 more frequently than T3, while tokens ending with relatively lower pitch were recognized

as T3. In the current investigation, we include offset pitch height as one of the three acoustic measures to determine the acquisitional stage of T2 and T3.

While many well-known perceptual studies (Liberman et al, 1957; Lisker and Abramson, 1970; Pisoni, 1973; Pisoni and Tash, 1974; Goldinger et al, 1996) have suggested that L1 perception of consonants and vowels tends to be categorical, McMurray et al (2002) used eye-tracking techniques and found that their participants were, in fact, sensitive to the fine-grained VOT difference in perceiving the English /p/-/b/ contrast. The categorical nature of phonemic perception shown in many studies was due to the categories that have been established earlier in life which put the speakers of a language into a "forced choice task" in most perceptual studies. However, it seems that humans can actually detect fine-grained acoustic differences. While many researchers agree that language perception is a categorical process, others assume that both L2 perception and production are naturally gradient processes. However, very few studies, except for the ones cited in the previous section of covert contrast, have shown a gradient nature of language acquisition, especially on a tonal level. Therefore, below, I state the big-picture question, in (1), and the hypothesis, in (2), that the study hopes to contribute to the field of language acquisition:

(1) Is L2 acquisition of phonemic categories a gradient or a categorical process?

(2) L2 acquisition is not a completely categorical process with respect to phonological contrasts.

It is the goal of this study to show that L2 acquisition of phonemic categories is not completely a categorical process, one in which an intermediate stage does not occur

between an initial stage where no contrast is implemented and a stage in which the production is target-like. Instead, it is a process in which an intermediate stage occurs between when the contrast is not being overtly produced and when the contrast is being fully implemented. Therefore, the specific hypothesis that is being tested in the present study is stated in (3):

(3) L2 learners of Mandarin may implement a measurable, statistically reliable distinction between T2 and T3 that is not perceived by native speakers of Mandarin.

Moreover, no longitudinal studies have been conducted to investigate the L2 acquisitional stages in terms of covert contrast and overt contrast in Mandarin. The current study aims to collect production data for a second time from some of the L2 participants eighteen months after the initial collection, and to conduct an analysis which will compare the data from the two time points. While the production data at one point will be able to test the hypothesis of the occurrence of covert contrast, a longitudinal analysis will allow testing whether a covert contrast is an intermediate stage of acquisition. That is, with longitudinal data, the study can examine whether any L2 participants made progress during the eighteen months between the first and second data collection. The hypothesis about the longitudinal analysis is stated in (4).

(4) Some L2 participants will progress over time from a stage of no contrast to implementing the T2-T3 contrast covertly, or from implementing the contrast covertly to one of implementing the contrast overtly.[1]

2.3 The Perception of T2 and T3

In addition to examining the intermediate stage in the T2-T3 acquisition process from production data, it is also the purpose of the current study to look at the process of acquiring T2 and T3 from the perspective of L2 perception.

Many studies, including Howie (1976) and Jongman et al (2006), have shown that F0 is the primary cue in perceiving Mandarin tones. Howie tested the tone perception in three pitch conditions with synthetic stimuli. One of the conditions was generated based on naturally produced F0 movement whereas the other two conditions had little to no information on the F0 patterns. It was found that the Mandarin native listeners performed reliably better in identifying the stimuli when the natural F0 information was available than when the little F0 pattern was included in the stimuli. This shows that the F0 pattern of tones is crucial in perceiving Mandarin tones.

Specifically, on one hand, according to Shen and Lin (1991) and Moore and Jongman (1997), the perception of T2 and T3 is determined by the F0 change from the contour onset to the turning point, which involves both F0 height and F0 contour, and the duration between the onset and the turning point. On the other hand, Blicher, Diehl

[1] The analysis of the data gathered at the initial time point does not reveal any participant with no contrast; all participants show either a covert contrast or an overt contrast. Therefore, I do not have the necessary data to test in the proposed longitudinal analysis the cases where a participant can move from no contrast to covert contrast, as proposed in Hypothesis 4).

and Cohen (1990) suggests that the overall syllable duration also affects the identification of T2 and T3 by both trained native speakers of Mandarin and English by conducting a study in which the targets with natural F0 contours were modified in the syllable duration. From those studies, it seems that both the F0 change between the onset and the turning point, the time that it takes to reach the turning point and the overall syllable length all function as factors in perceiving T2 and T3. Nevertheless, some studies (Moore and Jongman, 1997; Jongman et al, 2006) maintain the view that F0 typically serves as the predominant acoustic measure in perceiving and producing Mandarin T2 and T3.

2.4 L2 Perception and Production

In the field of L2 acquisition, there has been a large body of literature examining the relationship between production and perception of L2 sounds. There are several models that make proposals to explain the difficulties found in L2 learners' production and perception of TL contrast. Among them, the Perceptual Assimilation Model (PAM) and the Speech Learning Model (SLM) are the two of the most widely known and tested. PAM focuses on L2 learners' perceptual ability of a TL phonemic contrast and explains it on the basis of the way the L2 phonemes are assimilated to L1 phonemic categories. Specifically, PAM (Best, 1995; Best and Tyler, 2007) predicts that L2 learners will be able to discriminate two L2 sounds that are categorized into two different L1 phonemic categories better than two L2 sounds that are categorized as the same one L1 phonemic category. PAM also predicts that if L2 sounds cannot be categorized into any

L1 phonemic category, the perceptual accuracy will be from fair to good. Best, McRoberts and Goodell (2001) tested American English native speakers' perceptual performance of three Zulu contrasts, namely /ɬ/-/ɮ/ (two L2 phonemes categorized into two L1 categories), /k^h/-/k'/ (two L2 phonemes categorized as a good and a bad example of one L1 phoneme), and /b/-/ɓ/ (two L2 phonemes categorized as equal examples of one L1 phoneme). They found that the American English participants in their study discriminated /ɬ/-/ɮ/ better than /k^h/-/k'/, which was better discriminated than /b/-/ɓ/. This finding supported PAM. Since the model predicts that perceptual ability is associated with how L2 sounds assimilate to L1 phonemes, it also implies that the native language of the participants has an effect on the perceptual patterns of a TL contrast. However, PAM does not make any predictions on the perception of a L2 phonemic aspect that is not employed contrastively in the native language.

Hao (2012) tried to address this issue by testing the perception of Mandarin tones by L2 leaners whose native languages are Cantonese or American English. The L2 learners in his study had learned Mandarin in post-secondary classroom curriculum for an average of 2.68 years. The study found that the perceptual performances on all Mandarin tones did not statistically differ between the Cantonese speakers and the English speakers. The English speakers' accuracy rate in identifying Mandarin Tone 1 and Tone 4 (around 85% and 90%) was higher than the Cantonese speakers while the accuracy rate in identifying the other two tones, T2 and T3, did not differ reliably between the two speaker groups. Hao suggested that this is possibly due to the discrepancy between the Mandarin tone inventory and the Cantonese tone inventory.

There are six lexical tones in Cantonese; three level tones, two rising tones and a falling tone. It is fairly difficult to categorize Mandarin tones into Cantonese tone categories since some of them may have similar contours at different pitch range and some of them share similar registers but have different tone shape. This finding indicates that tones that are contrastively used in Cantonese did not help with the L2 perception of Mandarin tones. It does not necessarily support or oppose PAM with respect to L2 phonemic contrasts that are not employed in L1 phonology, but shows that L2 learners' perception on a phonemic distinction that is not used contrastively in the L1 can reach a certain degree of accuracy.

Another model, the Speech Learning Model (SLM) proposed by Flege (1987; 1995), emphasizes the L2 production of sounds. According to the SLM, whether a L2 learner can successfully learn a L2 vowel and produce it at a native-like level is determined by how that L2 vowel is classified into the L1 vowel categories. If the given L2 vowel is different from any L1 vowels, the L2 learner has a greater possibility of producing that L2 vowel more authentically than other L2 vowels that are more similar to L1 vowels and, therefore, can be classified into a L1 vowel category. Bohn and Flege (1997) investigated the SLM through the perception and production of English vowels /æ/ and /ɛ/ produced by German speakers, and found that their results support the claims of the SLM in that L2 learners of English can produce native-like English /æ/ when given sufficient exposure.

In addition to supporting the SLM, Bohn and Flege also investigated the relationship between production and perception. In their study, the results of an

identification task conducted with synthesized stimuli on the continuum of English /bæt/ and /bɛt/ was compared with the production of the two targets on two acoustic measures; vowel height and vowel length. They found that, on both of the measures, the perception and production of /æ/ and /ɛ/ were largely independent of each other. That is, for both the native control group and the two L2 speaker groups, no clear correlation was found between perception and production on vowel height and vowel length. However, one of the findings indicates a different interaction between perception and production for different acoustic measures. On one hand, when the L2 participants produced a large difference in vowel height between the two vowels, these participants also used vowel height as a primary cue to identify the two vowels. This suggests that there is, to a limited degree, a positive relationship between the perception and production on this specific acoustic measure of vowel height, even though no clear correlation was found. On the other hand, when some of the L2 participants produced a relatively large durational difference between the two vowels, these L2 participants were likely to make little use of the durational cue perceptually. In other words, there was a negative relationship between the produced contrast and perceptual cues in vowel length.

Moreover, Bohn and Flege discussed how being in different stages of the acquisitional process influences the interaction between perception and production. They recruited two L2-learner groups who had different amounts of L2 experience in English; the experienced German learners had been living in the US for more than five years whereas the inexperienced learners had only recently moved to the US, and had

been living in the US for an average only a little more than seven months. The results of this study showed that, for the inexperienced L2 group, the learners could sufficiently perceive the distinction between /æ/ and /ɛ/, however, they could not systematically produce the contrast. On the other hand, the experienced L2 group showed a tendency that the perception of /æ/ and /ɛ/ falls behind production. This finding indicates that the current stage of a L2 learner in the acquisition process has an effect on how perception and production interact with each other.

These findings of Bohn and Flege are particular interesting for the present study in which one of the goals is to test the perceptual ability of two different groups of L2 learners. The two groups of learners in the current study showed the T2-T3 contrast in Mandarin differently in terms of whether the contrast is perceived overtly by the native speakers of Mandarin. In the present study, it is also of the interest to discuss the relationship between perception and production of L2 learners at different stages in producing T2 and T3 in Mandarin.

Similar to Bohn and Flege regarding the relationship between perception and production, Peperkamp and Bouchon (2011) examined the acquisition of English /i/ and /ɪ/ produced and perceived by L2 speakers whose native language was French, and found no correlation between the production and perception of the target phonemes. Their study used two measures to access production. Native-speaker judges rated the productions of the vowel, /i/ and /ɪ/, in terms of accuracy, i.e. whether or not the judges perceived the contrast, and in terms of the how native-like the productions were. In addition, this study used an ABX test to measure the L2 participants' perception. One

of the findings is that the error rates of the ABX test were not correlated with either of the assessment measures of the production test. Peperkamp and Bouchon accounted for this lack of correlation between production and perception in terms of the short inter-stimulus interval used in the ABX test. According to the authors, during the perception task, a short inter-stimulus interval hindered the L2 participants' access to the phonological information that was used in production. Therefore, the authors suggested that the way perception and production interact with each other in L2 acquisition largely depends on how much information on L2 phonological categorization the L2 learners can access during the perception task.

Contrary to the studies that show no clear relationship between perception and production, some other studies (Ingram and Park, 1997; Bradlow, Pisoni, Akahane-Yamada, and Tohkura, 1997) suggest a correlation between these two aspects of L2 acquisition. For example, Ingram and Park (1997) examined the L2 production and perception of Australian English vowels by native speakers of Japanese and Korean. The vowels in Japanese differ from those in Australian English in that Japanese lacks /æ/, and by virtue of the fact that Japanese vowels evidence a duration contrast. Korean vowels differ from those of Australian English in that Korean lacks /æ/ and /e/, however, there is no duration contrast in Korean as there is in Japanese. Ingram and Park found that the Japanese participants identified the targets reliably better than the Korean participants, especially on /æ/ and /e/. The acoustic analysis of /æ/ and /e/ produced by the Japanese participants indicated a clear separation between the two targets in vowel quality plotted by measuring the first formant and the second formant, whereas

the Korean participants produced these two vowel targets with an overlap between the first and second formant. Altogether, the Japanese participants produced and perceived /æ/ and /e/ systematically better than did the Korean participants. This suggested that the participants' production mirrors their perception when the participants are on the basis of their L1, which, in turn, indicates that there is not a weak positive correlation for the groups between the perception and the production of the vowels.

Bradlow et al (1997) obtained a finding that is similar to Ingram and Park (1997) based on the L2 perception and production of English /r/ and /l/ by Japanese speakers. Different from Ingram and Park, which collected one-time data, Bradlow et al. with data gathered at two time points focused on investigating the effect of perceptual training in both perception and production ability. The authors conducted a pretest and a post-test, both of which included a perception task and production task. The perception tests were simple identification tasks in which the L2 learners listened to the production of native speakers of American English and identified whether they heard /r/ or /l/. The participants' production ability was determined using two different judgment tasks. The first was a preference task in which native speakers of English listened to the same tokens produced in the pre-test and the post-test, and then picked a "better" representation of the target English word shown on the screen. The second test was a simple identification task in which the native speakers transcribed the L2 production of /r/ and /l/. In order to investigate the effect of perceptual training on the participants' L2 production and perception, the participants were given, between the pre-test and the

post-test, a lengthy training period consisting of 45 sessions. The results showed that the perceptual training given to the L2 participants between the pre-test and the post-test helped improve not only the perceptual but also the production performance. In addition, a positive correlation between the overall accuracy in perceptual identification and the overall accuracy of native speakers' identification of the L2 learners' production was found regardless of some individual variations. That is, a high accuracy in one indicates that the L2 participant is likely to reach a high accuracy in the other aspect. In addition, the central finding from Bradlow et al. that perceptual training actually helped improve the L2 participants' production of /r/ and /l/ suggests that there is a link between L2 perception and production. Instead of being independent from each other, this study suggests that the perception and production of a non-native contrast share a common mental representation that operates integrally on the tasks of perception and production.

One point that is worth noting is that although Bradlow et al. found a positive relationship between the overall accuracy of perception and production, no clear relationship was found between the improvement of perception and production from the pretest to the post-test. The L2 participants showed individually different rates of improvement on identifying and producing the targets. This suggests that more data need to be gathered in order to decide how improvement on one aspect helps the other improve. One plausible way to approach this issue is to examine L2 production and perception from a different perspective, instead of using the classic identification tasks which are categorical in nature. As summarized earlier in the section, it has been

shown that language users sometimes have the ability to detect a fine-grained difference on an acoustic measure that is relevant in making a contrast in their languages (McMurray et al, 2002). Also, studies in both L1 and L2 have suggested that language learners can at times produce a statistically reliable contrast that is not necessarily perceived by the native speakers of the TL. Perhaps, whether or how L2 learners can perceive the fine-grained acoustic difference on the acoustic measure relevant to a L2 contrast can provide new information on L2 perception and its relationship to production in general. Although no clear hypothesis can be proposed and tested in terms of the relationship between improvement in perception and production (as raised by Bradlow et al), a production test which aims to examine a covert contrast within a longitudinal analysis can possibly shed some light on the interaction of the improvement between perception and production.

Several other studies (Flege and Schmidt, 1995; Flege, MacKey, and Meador, 1999; Kluge, Rauber and Reis, 2007) have also suggested a positive relationship between L2 production and perception. Flege and Schmidt (1995) investigated the effect of speech rate on the L2 acquisition of the English aspirated voiceless bilabial stop [p^h] by native Spanish speakers. English and Spanish voiceless bilabial stops differ in their VOT in that Spanish /p/ is produced with short-lag VOT (roughly from 0 to 25 milliseconds between the release of the plosive and the onset of voicing) whereas the English /p/ (i.e., the allophone [p^h]) is produced with a long-lag VOT (roughly 60 to 100 milliseconds between release of the closure and the onset of voicing). In the Flege & Schmidt study, L2 participants were divided into two groups labeled “proficient” and

"non-proficient" by a group of English native speakers, based on the amount of foreign accent that was detected on the basis of their L2 production of some stimulus sentences. The participants then listened to several examples of synthesized /p/ stimuli which differed in VOT values. The synthesized stops occurred in syllables of two durations, 125 and 325 milliseconds. The participants were asked to give a rating to each stimulus in terms of how representative it was of English [p^h]. Schmidt and Flege showed in this study that the long lag allophone of English /p/ shows more variation than the short lag /p/ of Spanish as a function of the speech rate. Therefore, it was expected that the perception of English /p/ by the monolingual Spanish speakers would be less affected by the speech rate. The authors found that the goodness ratings of synthesized stimuli with different VOT values of /p/ that were assigned by both the native group and the proficient L2 group were significantly affected by the syllable duration, but the ratings given by the non-proficient L2 group showed no effect of syllable duration. In other words, the proficient group of Spanish speakers resembled the English native speakers group in the effect of speech rate on the perception of English /p/ while the non-proficient group showed little resemblance to the native speakers of English. This finding indicates a somewhat positive correlation between the Spanish speakers' production of English and their perceptual rating of English /p/. As the production of the proficient L2 learners was judged as having less trace of foreign accent than the non-proficient L2 learners, their perceptual goodness rating resembled the native speakers of English more than the non-proficient group.

The method that Flege and Schmitd used to assess their L2 learners' proficiency in production has been widely used in the field of L2 acquisition. It is plausible to evaluate L2 production based on the degree of foreign accents rated by the native speakers of the TL, since the judgment of these native speakers serves as the most straightforward method to determine how native-like the production of a L2 learner is. In addition, such an approach is presumably one of the only few ways to assess L2 learners' natural production in the TL. The current study will also use a similar method, native speakers' transcription, as one of the ways to investigate L2 production. However, this method, which requires native speakers to give ratings on a scale of the amount of foreign accent or to transcribe L2 production, seems like a relatively subjective approach, which might be affected by other factors, including the native speakers' language exposure and language preferences. Given that the rating of L2 production by TL native speakers can be subjective, it is not always reliable to depend solely on this approach when assessing L2 learners' productions. Therefore in this study, L2 learners' productions will be analyzed for the implementation of a covert contrast, not only on the basis of native speakers' transcriptions, but also in terms of acoustic measurements.

As there are studies on both sides of the debate over whether there is a clear relationship between L2 perception and production, the present study hopes to contribute to the discussion by examining production data from the point of view of a covert contrast, and by analyzing perception data using the production results as a criterion to group L2 participants. Last but not least, no studies have investigated the

relationship between L2 perception and production from the perspective of a stage of L2 acquisition in which a learner produces a statistically reliable contrast that is not overly perceived by the native speakers. Nor have any studies looked at this stage of covert contrast in the acquisition of Mandarin T2 and T3. The significance of this study lies in providing insights into the ongoing debate on L2 perception and production from an approach that assesses L2 learners' production acoustically.

2.5 Hypotheses

The hypotheses that the current study aims to test are stated in this section. First, the hypothesis for the production test is listed here again for convenience as (3):

(3) L2 learners of Mandarin may implement a statistically reliable acoustic distinction between T2 and T3 that is not perceived by native speakers of Mandarin.

In addition to (3), as mentioned previously in this section, I also gathered longitudinal data from some of the L2 participants. The hypothesis that the longitudinal analysis will test is repeated as (4).

(4) Some L2 participants will progress from not showing any contrast to implementing the T2-T3 contrast covertly, or from implementing the T2-T3 contrast covertly to implementing the contrast overtly.

Based on (3), the present study is also interested in how the perception differs between the learners who produce an overt contrast and the learners who produce a covert contrast between the target tones. In order to address this question for T2 and

T3, two perception tasks have been carried out: an identification task and a discrimination task.

In the identification task, the L2 participants were asked to transcribe the targets which were naturally produced by a native speaker of Mandarin. The relevant hypothesis is stated as in (5).

(5) The L2 participants who produce an overt contrast between T2 and T3 will identify the T2 and T3 statistically better than the L2 participants who implement the contrast covertly.

In the discrimination task, the L2 participants were asked to listen to a pair of stimuli that were pitch-modified at the turning point and the contour offset, and to decide if the stimuli of the pair were the same or different. The discrimination task was conducted to test the hypotheses stated in (6) and (7):

(6) The performance on the discrimination task by L2 participants who produce an overt contrast between T2 and T3 on the production task will not be reliably different from the performance on the discrimination task of the native speakers of Mandarin.

(7) The performance of L2 participants who implement a covert contrast will be systematically different from that of the native speakers of Mandarin.

The rationale for (6) and (7) is the expectation that the L2 participants who produce an overt contrast will show results that approximate that of the native speakers since their L2 production of T2 and T3 has been overtly perceived as target-like. I except to the L2

overt participants demonstrate a systematically different result on the discrimination task from the participants who produce a covert contrast.

This study consists of one production and two perception tasks. The overall design of the study is in four parts: 1) L2 learners were recorded producing T2 and T3 targets which were later analyzed acoustically to see if they have been implementing a statistically reliable distinction between the two tones; 2) in the longitudinal analysis, the results have been compared with data that have been collected from some of the current L2 participants who participated in a previous study of their production of T2 and T3; 3) L2 learners were tested on an identification task which consisted of targets containing T2 and T3 produced by native speakers, in which data were analyzed in conjunction with the production data to see if L2 learners whose production was transcribed as target-like by native speakers also perceive T2 and T3 statistically better than learners who produce the contrast covertly. And finally, 4) a discrimination task with pitch-modified targets was conducted to both groups of the L2 participants, native speakers of Mandarin and American English speakers with no exposure to Mandarin tone, to investigate if there is any a reliable difference among the results of the different participant groups.

In the next three chapters, the tasks are presented in sequence. Chapter 3 concerns the production task, reporting the methods of data collection and the analyses of its results, including the longitudinal analysis. Chapter 4 describes the methods and findings of the identification task. Details of the discrimination task and its results are presented in Chapter 5. The methods of all the tasks are reported in terms of the details

of the participants, stimuli, procedures, acoustic analyses and the approaches of the statistical analyses.

Chapter 3

Production

This chapter reports the methodology of the production data and findings of the investigating the occurrence of an intermediate stage in L2 learners' acquiring the T2-T3 contrasts in Mandarin. First, in Section 3.1, the methods that were used to obtain the production data are described in details. Secondly, Section 3.2 and 3.3 each reports the analytic details and findings of Analysis 1 and Analysis 2. Note that the production data were gathered during two elicitations in order to perform a longitudinal analysis. The data gathered at each time point were analyzed twice, namely according to the criteria of Analysis 1 and Analysis 2, which are discussed in detail later in the chapter. Finally, the last section of the chapter 3.4 discusses the findings of the longitudinal analysis, concerning whether the L2 speakers made a longitudinal progress from the first elicitation to the second elicitation.

3.1 Methodology

3.1.1 Participants. Tonal productions were elicited from sixteen L2 learners of Mandarin, all of whom have studied Mandarin at the same university in the Midwest for at least two years. Among the sixteen L2 participants, six have participated in the production tasks conducted at both points in time for the longitudinal analysis. At the first elicitation, four other L2 learners participated, for a total of ten L2 participants. In the eliciting production data at the second time points, six new participants were recruited, which gave total of twelve L2 learners at the second time. All of the L2 participants are native speakers of American English. They were undergraduate students at the time of

the task; seven are females and nine are males. None of them has reported any disorder related to speaking, listening or learning. All L2 participants can read tone marks and pinyin, which is the alphabetic spelling system used in the field of teaching Chinese as a second language. They are not linguistically trained and were not informed of the purpose of the study.

The L2 participants have been assessed by the Oral Proficiency Interview (OPI) guidelines of the American Council of the Teaching of Foreign Languages (ACTFL). OPI provides a standardized guideline to evaluate the global functional speaking proficiency of foreign language learners. It is also the most used and recognized proficiency assessment in foreign language curriculum in the United States. During an OPI test, an ACTFL certified OPI tester interviews a foreign language learner in a face-to-face or telephonic setting. There are usually multiple communication tasks in which the tester compares a leaner's speaking ability with others' at different proficiency levels on the guidelines. Although the primary goal of the OPI is to assess a learner's functional oral proficiency, instead of a non-native accent, a learner's nativeness and accuracy of pronunciation also affects one's functional ability and listeners' understanding of the non-native speech. There are eleven proficiency levels on the OPI guideline; distinguished, superior, advanced-high, advance-mid, advanced-low, intermediate-high, intermediate-mid, intermediate-low, novice-high, novice-mid, and novice-low. An OPI tester at a Midwestern university ranked the L2 participants in the study between Intermediate-mid to Advanced-mid, suggesting that their Mandarin proficiency is at or above the intermediate level. Learners at this level were chosen to participate in the

study because it is assumed a stage of covert contrast is most likely to be attested by learners at an intermediate level or higher. Moreover, as mentioned, six of the L2 participants have participated in the production task at the first elicitation, which was seventeen months prior to the second elicitation. From the production data gathered at the first time, the preliminary results support this assumption that learners at an intermediate level or higher show a tendency of implementing the T2-T3 contrast covertly. This design allows a longitudinal analysis of the data from the six L2 participants who participated at both time points, and may show their progress from one stage to another in the acquisition of the contrast.

Two native speakers of Mandarin participated as the judges to determine whether they perceived the participants' productions of the tones as target-like. Both of the judges are native speakers of Mandarin from Taiwan.

3.1.2 Stimuli. Identical stimuli were used in both of the elicitations, which are all monosyllabic words with a CV phonetic structure. Both elicitations used only CV structures because nasals, which are the only consonant category that is allowed in the coda position in Mandarin, are known to affect the vowel quality and fundamental frequency in many world languages, including Mandarin (Ohala, 1975; Chen, 2000). The vowel in the CV syllable was always one of the followings: /i/, /a/, /u/, /ü/ and /ʌ/. Monosyllabic words were chosen in order to avoid any tonal effect carried from neighboring tones in context. The stimuli list consisted of 39 minimal pairs of the T2-T3 contrast, 78 targets in total with 78 fillers, which are T1 and T4 monosyllabic and some disyllabic words. In total, 156 words were produced by each participant at both

elicitations and the six participants who participated in both of the elicitations produced the same list of words.

Two randomized lists were created for the L2 speakers to read. There were ten practice items on each list before the start of the target word list. None of the practice items were on the target list. All stimuli were typed in pinyin with tone marks in order to elicit the best possible pronunciation, as Chinese characters do not provide any tonal information. Tone marking in Mandarin, in fact, indicates the contour of the pitch of the tone, respectively *mā* for T1 (high level), *má* for T2 (rising), *mǎ* for T3 (dip-rise) and *mà* for T4 (falling). Therefore, tonal information was given to the speakers as much as possible for both production tasks in order to collect the data with learners' best and most careful pronunciation.

3.1.3 Procedures. All of the L2 speakers were recorded individually in a sound proof recording booth or a quiet room using the computer program, *Audacity*. During the recording the L2 speakers were asked to read one of the two lists and to pronounce the items on the list with a normal speech rate. They were instructed to read the practice items before reading the words on target list. They were also requested to pause at least one second between each item in order to avoid context effect that would affect tone contour.

After collecting all the recordings, noise reduction was conducted on Audacity. Individual words were then cut out from the recordings and organized using *Praat*. In the task involving native speakers' judgment, the individual sound files of targets were played to the native listeners. The listeners were given answer sheets numbered with

the items prior to the task and requested to write down what tone they think they hear intuitively without thinking about what the speaker had intended to say. The data of the native speakers' judgment were entered and organized in an Excel file.

3.1.4 Analysis: The data were analyzed according to two different methods, leading to two analyses. The reason for conducting two analyses is that if covert contrast can be attested through different analytical methods, we thereby can be more certain of our documentation of this intermediate stage. The two analyses differ in the following two ways. First, the criterion for scoring a token as target-like based on the transcription results is different. Below I will describe the details and the reasons behind using different criteria. Second, a criterial threshold was established in Analysis 2 for L2 learners' target-like production. In Analysis 1, there was no criterial threshold set for whether a participant has produced the contrast according to the transcribers. The data were analyzed based on whether L2 targets are overtly perceived by the native transcribers or not. In Analysis 2, L2 participants were credited with producing the contrast if their production reaches the criterial threshold of being target-like.

In the following sections, the methods are reported for each of the corresponding analyses; factors concerning the acoustic and statistical analysis and the criteria of scoring the transcription are detailed before presenting the results in each analysis. In each analysis, the results of the two elicitations are reported separately, beginning with Analysis 1. Lastly, a longitudinal comparison of Analysis 1 is presented before proceeding to Analysis 2.

3.2 Analysis 1

3.2.1 Analysis Method

Variables. The predictor conditions, sometimes known as the independent variables, are the input, which in this study are the T2 and T3 targets. Four outcome variables (i.e. the dependent variables), the output, were investigated in the present study. Specifically, they are the percentage of tones being overtly perceived by the native-speaking judges, the differences between the F0 frequencies, the duration between the contour onset and the turning point, and lastly, the offset pitch height.

Native listeners' transcription. A covert contrast is defined as a statistically reliable acoustic distinction that is not perceived by native speakers of the TL. The criterion for scoring a token as target-like was initially that at least one of the native transcribers heard it as intended. However, in an analysis conducted under this criterion no statistically reliable difference was found, either for the group results or the individual scores. Therefore, another analysis was conducted using a stricter criterion: a target is considered target-like when both native transcribers heard it as target-like. The results presented in Analysis 1 were obtained under the stricter criterion.

Acoustic analysis. The acoustic measures were analyzed using *Praat*. As described above in the section on outcome variables, three acoustic measures were used: F0 difference, the duration to the turning point and the offset pitch height. The turning point is defined as the point in time where the F0/pitch stops falling and starts rising. In order to get the F0 difference, the lowest F0 in the contour and the highest F0 before the turning point were both measured using *Praat*. F0 difference was

obtained by subtracting the lowest value from that of the highest F0 before the turning point. The point where the lowest F0 occurs is the turning point. *Praat* scripts were used to obtain these acoustic measures along with the offset pitch height. The F0/pitch setting were set differently for male and female participants. Pitch range was set from 75 to 300 Hz for males and from 100 to 500 Hz for females. The measurement of the duration to the turning point was normalized with the total duration of the pitch contour shown on the spectrograms. For most targets, pitch contour appears only in vowel segments; therefore, the total duration of the pitch contour approximates the vowel length. However, for targets whose onset consonant is a nasal, the onset of the pitch contour was measured at the nasal onset, given that nasals carry a pitch themselves, and it is difficult to separate the onset of the nasal from the nasalized vowel.

Statistical analysis. In order to determine the implementation of a covert contrast, paired-sample t-tests that compare the means of T2 and T3 would ideally be conducted to determine whether the distinction in the acoustic measures between T2 and T3 is statistically significant. However, the sample sizes of NT2 and NT3 are different for each participant, as are the sample sizes of the overtly perceived tokens of T2 and T3, and paired-sample t-tests cannot be used without compromising some data points. Therefore, a linear model (sometimes referred as regression), which enables an analysis with uneven sample sizes, were used to determine whether each of the four outcome variables was a function of target tone. In other words, a linear model tests whether an outcome variable was predicted by the target tone. Linear model tests were also used to analyze whether a variable was reliably predicted by all the intended target

tones and the tones that were overtly identified and perceived. A 95% confidence interval is set for this study to determine the significance of a respective statistic model.

In the longitudinal analysis, which investigated if there is a change in the native transcription between the two elicitations and if there is any interaction between the tones and the elicitations, two statistical methods were used. I used McNemar's test to determine if the native transcriptions show a significant change for each participant's target-like productions. Also, a linear model with two predictor variables, tones and elicitation, were employed to see if there's an interaction between tones and elicitations.

The results will be discussed in view of the production hypothesis, as stated in the previous chapter, namely, that L2 learners of Mandarin will implement the contrast between T2 and T3 in a way that is not perceived by native speakers of Mandarin. In testing this hypothesis, we will use the results from individual participants rather than aggregated data, because we are asking whether the tones in question have been acquired. In so doing we are inquiring as to the state of a participant's interlanguage grammar. An interlanguage grammar is constructed individually in every L2 learner's mind, and as such must be viewed individually. Therefore, only individual results were presented for the discussion of L2 production.

3.2.2 Results: First Elicitation

This section reports the data of each L2 participant from the first elicitation. I will first present the results of the native speakers' transcription on the L2 productions of

tones, followed by the results of the statistic analysis of the target tones that were not transcribed as target-like and the ones that were overtly perceived.

Native Transcription. For each individual L2 learner, the percentage of the targets that were overly perceived was presented in Figure 3.1. There are three points that are worth noticing from the figure. First, every participant showed a lower percentage of T3 being overly perceived than T2. Secondly, eight of the participants produced at least 80% of the target-like T2 while only one participant produced target-like T3 on more than 50% of the attempts. The first two points are consistent with those of some of the studies on T2 and T3 (Shen, 1989; Elliott, 1991) as well as my own observation in L2 Mandarin instruction. That is, it is more difficult to produce target-like T3s than to produce target-like T2s for most Mandarin L2 learners. Most of the T3 that were not perceived as target-like were transcribed as T2 by the native transcribers, which provides an ideal context for investigating whether there is a statistically reliable difference between the T2 and the non-target-like T3. Lastly, all participants showed a great difference between the percentage of target-like T2 and that of target-like T3, especially Participant BHR and Participant LKL. All of the T2 Participant BHR produced were transcribed as target-like, but only 5% of his T3 attempts were perceived as target-like. A similar situation occurred with Participant LKL.

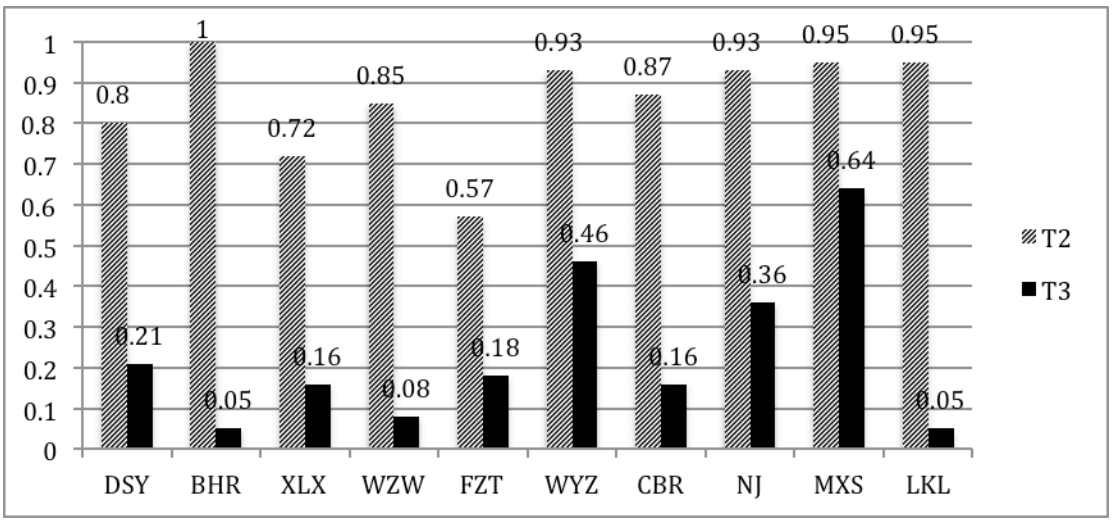


Figure 3.1 Percentage of targets that were overtly perceived by the native transcribers (First elicitation: November, 2013)

Statistical Analysis. This section reports the statistical results of individual participants for each variable. Each participant's data was analyzed independently from each other, and a table that reports the statistics details was created for each participant (Appendix A). Table 3.1 is an example of the statistics details of Participant DSY's production, illustrating how the statistics results were presented in detail for each participant. For the convenience, a summary table (Table 3.2) was created here for presenting the significance of the results. It summarized the statistical results of each of the acoustic variables measured from each L2 participant's production.

In Table 3.1, the first column lists the three acoustic variables tested as functions of the target group pairs, listed in the next column. The three pairs are 1) NT2 and NT3, 2) Overt T2 and Overt T3, and 3) all T2 and all T3. The third column from the left is the mean value of the acoustic variable measured for the target group. And lastly, the last column reports the significance value of the linear model constructed for the respective target pair. Cells were shaded to indicate a significant p value. For example, from the

first two rows below the header row, the linear model of F0 difference as a function of target tone for the pair of NT2 and NT3 is not significant, with a p value of 0.9. However, I also conducted a model of duration to the turning point as a function of NT2 and NT3 (the first two rows of the *variable duration to the turning point*), and this model is significant *(t(40)=2.302, p = 0.02, p<.05*).

Table 3.1 Participant DSY: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3

Variable	Pair	Mean	Significance test
F0 Difference (Hz)	NT2	39.4	t=0.115, p=0.9, p>.05
	NT3	41.4	
	Overt T2	20.9	t=2.362, p=0.024, p<.05
	Overt T3	39.4	
	All T2	24.7	t=2.08, p=0.04, p<.05
	All T3	41.1	
Duration to the Turning Point	NT2	0.23742	t=2.302, p=0.02, p<.05
	NT3	0.32778	
	Overt T2	0.19913	t=2.764, p=0.009, p<.05
	Overt T3	0.35564	
	All T2	0.20699	t=5.1, p=0, p<.05
	All T3	0.33135	
Offset Pitch(Hz)	NT2	291.05	t=-1.545, p=0.13, p>.05
	NT3	275.55	
	Overt T2	306.98	t=-3.3, p=0.002, p<.05
	Overt T3	262.87	
	All T2	303.71	t=-4.9, p=0, p<.05
	All T3	273.92	

To sum up the details from Table 3.1, for the NT2 and NT3 produced by Participant DSY, only the duration to the turning point was reliably predicted by the target tones, but not the other two variables. For her overt T2 and overt T3, all the three variables were statistically reliable as functions of the tones, so as for the pair of all T2 and all T3. The description above is the summary of Participant DSY's data, and a summary for each participant was transformed to a single row in Table 3.2. For the detailed results of the statistics test for other nine participants, please refer to Appendix

A, which provides nine independent tables that presents the results in the same fashion as Table 3.1.

In Table 3.2, each row summarizes the significance of the linear model constructed for the three target pairs on the three variables. The first column at the left side is the participant (DSY, BHR…to LKL). The second header row from the top is the pair of targets, respectively NT2 and NT3, Overt T2 and Overt T3, and All T2 and T3. The row below the target pair lists the three variables for each pair. Finally, the results of the significance test were presented by plus marks, indicating significance, and minus marks, showing insignificant models. In other words, a plus mark shows that the linear model of the variable as a function of target tones is significantly reliable, whereas a minus mark indicates that there is no significance found for the linear model.

Table 3.2	**Acoustic Measures: Are the variables reliably predicted by the target tones?**								
Pair	**NT2 vs NT3**			**Overt T2 vs Overt T3**			**All T2 vs All T3**		
Participant	F0 Diff	Duration	Offset	F0 Diff	Duration	Offset	F0 Diff	Duration	Offset
DSY	-	+	-	+	+	+	+	+	+
BHR	N/A	N/A	N/A	+	+	+	-	+	+
XLX	-	-	-	+	+	+	+	+	-
WZW	-	-	-	+	+	+	+	+	+
FZT	+	+	-	+	+	+	+	+	+
WYZ	-	-	-	+	-	+	+	-	-
CBR	-	+	+	+	+	+	+	+	+
NJ	-	-	-	+	+	-	+	+	-
MXS	-	+	-	+	+	+	+	+	+
LKL	-	-	-	+	-	+	+	+	+

Before discussing the general findings of the participants, note that there is no data shown for the pair of NT2 and NT3 produced by Participant BHR. This is because that all of his T2 productions were transcribed overtly by the native transcribers, which can be seen from Figure 3.1 in the previous section.

Let us now proceed to discuss the L2 participants' statistic results on the acoustic variables. First, we look at the NT2 and NT3 pair. Any plus marks in the columns of NT2 and NT3 would support the hypothesis of the occurrence of an intermediate stage of covert contrast in acquiring the L2 Mandarin T2-T3 contrast. Four (DSY, FZT, CBR and MXS) of the ten participants showed at least one acoustic variable significantly predicted by the target tone even though the tones were not transcribed as target-like. It is shown that all four of them made the contrast in the duration variable, which is a statistically reliable function of target tone. In addition to the duration variable, Participant FZT also implemented the contrast in the F0 difference, as it is reliably predicted by the contrast. Participant CBR's result shows that his NT2 and NT3 also significantly predict the offset pitch. The other five participants' acoustic measurements do not indicate any significance in being a function of the contrast in cases where the target tones were not transcribed as target-like.

We now look at the columns under the Pair Overt T2 and Overt T3. These columns show the statistic results of the acoustic variables of the targets that were transcribed as target-like by the native transcribers. First, in this target pair, the results show, for all participants, a statistically reliable model of the F0 difference as a function of the contrast. That is, the contrast, whether the target is T2 or T3, significantly predicts

the F0 difference between the contour onset and the turning point. All participants produced significantly larger F0 drop between the contour onset and the turning point. Second, in addition to F0 difference, eight of the participants also produced a significant distinction in the duration variable that was predicted by the overly-perceived T2 and T3. The overt targets produced by Participant WYZ and LKL do not show significance in the duration variable. In terms of the last acoustic measures for the overt targets, nine participants produced offset pitch height that is significantly predictable according to the contrast, Overt T2 and Overt T3. Participant NJ's overt targets do not indicate any significant linear relationship between the target and the offset pitch.

To sum up the statistical results of the overt targets, seven of the participants produced overt targets that significantly predict all three acoustic variables. For all participants, F0 difference serves as a linear function of the overt targets. This finding preliminarily suggests that in order for the targets to be transcribed as target-like, a L2 participant needs to produce F0 difference that is linearly predicted by the targets, whether it is a T2 or T3. Also, the finding that three of the participants show a linear prediction of only two acoustic variables predicted by the overt targets might suggest that it is not required to have all three acoustic variables significantly predicted by the contrast. From the three participants (WYZ, NJ, and LKL), producing tones whose two acoustic variables are statistically reliable functions of the targets is sufficient.

The next section reports the data of individual L2 participants from the second elicitation, which was seventeen months after the first elicitation. As described in the methodology section, not all ten L2 participants were recorded for the second elicitation.

Six of them participated again, with another six newly-recruited L2 learners. The results of the second elicitation will be presented in the same way of the first elicitation, with the native transcription presented first, followed by the statistical results of the acoustic variables for each target pairs.

3.2.3 Results: Second Elicitation

Native Transcription. Figure 3.2 illustrates the results of the native transcription for each L2 participants in the percentage of targets being transcribed as target-like. The results of the six participants who have participated in both elicitations were displayed at the right-hand side of the figure, in the same order as shown in Figure 3.1 of the first elicitation. Identical to Figure 3.1, the striped bars are the T2 target-like percentage and the solid bars are the T3 target-like percentage. Among twelve L2 participants, ten showed similar results to the first elicitation in terms of which tone is more likely to be transcribed as target-like than the other one. Those ten participants produced more T2 than T3 that were transcribed as target-like by the native transcribers, which conforms with the result of the first elicitation and the overall observation in L2 Mandarin acquisition. Two participants' data showed the opposite. From Participant LJ and NJ's production, more T3 were transcribed as target-like than T2. For LJ, most (97.5%) of the targets he produced were perceived as target-like, 95% for T2 and 100% for T3 respectively. For NJ, 72% of the T2 and 85% of the T3 she produced were transcribed as target-like. Participant FMY did not produce any T3 that was transcribed

as target-like as there is no solid bar visible from the figure in her column. Therefore, there will be no statistical analysis conducted for her overt targets.

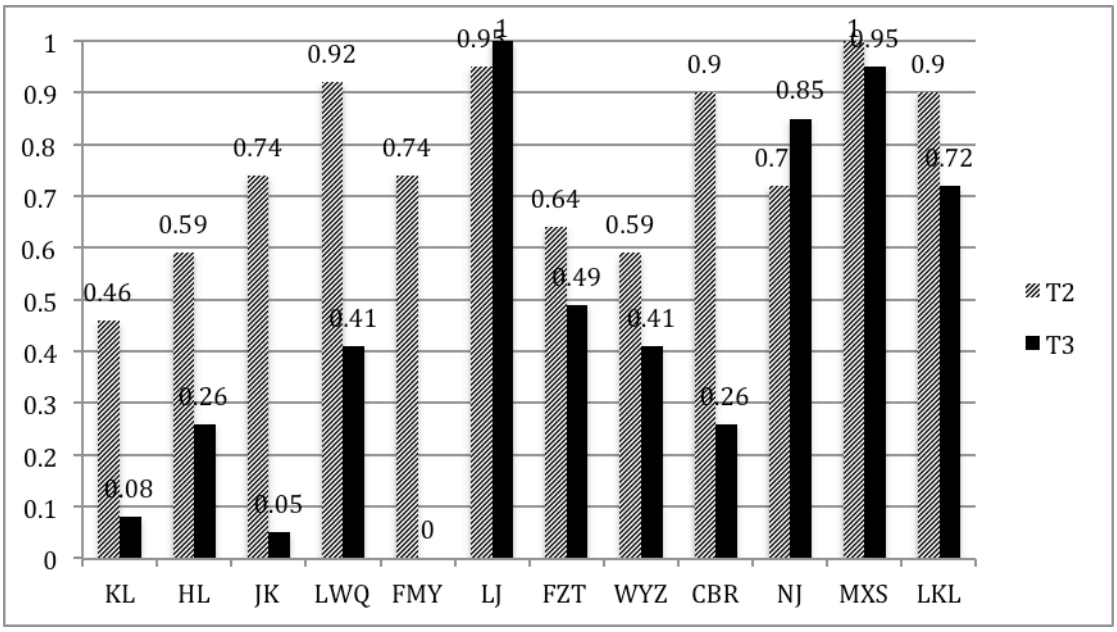


Figure 3.2 Percentage of targets that were overtly perceived by the native transcribers (Second elicitation: April 2015)

Also, compared to the first elicitation, there were more participants producing target-like T3 over 50% of the time. There were four participants (LJ, NJ, MXS and LKL), as opposed to only one in the first elicitation, and three of them participated in both of the elicitations.

Statistical Results. The result of each individual participant was organized in the same way as in the first elicitation, as shown in Table 3.1 in the previous section. All of the detailed individual results were attached in Appendix B. The individual results were summarized in Table 3.3 below, presented and organized in the same way as in Table 3.2.

Table 3.3	**Second Elicitation** **Acoustic Measures: Are the variables reliably predicted by the target tones?**								
Pair	**NT2 vs NT3**			**Overt T2 vs Overt T3**			**All T2 vs All T3**		
Participant	F0 Diff	Duration	Offset	F0 Diff	Duration	Offset	F0 Diff	Duration	Offset
KL	_	_	_	+	_	_	_	_	_
HL	_	_	_	+	_	_	+	_	_
JK	_	_	_	+	_	_	_	+	_
LWQ	_	_	_	+	+	_	+	+	_
FMY	+	_	_	N/A	N/A	N/A	+	+	_
LJ	N/A	N/A	N/A	+	+	+	+	+	+
FZT	+	_	+	+	+	+	+	_	+
WYZ	_	_	_	+	+	_	+	+	_
CBR	_	_	_	+	+	_	+	+	+
NJ	_	_	_	+	+	+	+	+	+
MXS	N/A	N/A	N/A	+	+	+	+	+	+
LKL	+	_	+	+	+	+	+	_	+

First, we look at the results of the linear models conducted individually for each participants' NT2 and NT3. Note that MXS and LJ have no data for NT2 and NT3 because MXS's T2 and LJ's T3 were all transcribed as target-like. Among the ten participants who produced targets that were not transcribed as target-like, three of them (FMY, FZT, and LKL) showed at least one acoustic variable significantly predicted by the target tone even though the tones were not transcribed as target-like. All of the three participants showed a linear relationship between F0 difference and the targets. Two (FZT and LKL) of the three participants showed a significant relationship in which the offset pitch was predicted by the targets. No statistically reliable relationship has been found between the duration and the targets from any participants' production of NT2 and

NT3. This is particularly interesting because from the first elicitation, among the three acoustic variables, duration is the one that shows being significantly predictable by the targets for more participants (four participants) than the other two variables (one participant each).

To sum up the results of the linear models of the acoustic variables as functions of NT2 and NT3, two participants made a statistically reliable contrast on two of the acoustic measures and one made a significant distinction between NT2 and NT3.

Next, we turn to the overt targets. As mentioned in the last section, Participant FMY did not produce any T3 that were transcribed as target-like by both of the transcribers. Therefore, there is no statistical data shown for her overt targets. The results show that all eleven participants produced overt T2 and T3 that significantly predict the F0 difference between the contour onset and the turning point. This complies with one of the results of the overt tones gathered from the first elicitation, which finds that in order for the production to be transcribed as target-like, the F0 difference is required to be a statistically reliable function of the target tones. In addition, among the eleven participants, eight produced the duration as a statistically reliable function of the overt T2 and overt T3. And five of them were found to have the offset pitch significantly predicted by the targets. Three participants, KL, HL and JK's overt tones show only one acoustic variable in F0 difference as the function of the contrast. Three participants, LWQ, CBR and WYZ, showed two acoustic measures (F0 difference and duration) reliably predicted by the contrast. Lastly, five made a statistically reliable distinction on all three acoustic variables between the overt T2 and overt T3.

We now compare the overt and the covert tones in terms of how the significance of the acoustic variables as functions of the target affects the target-likeness. From Figure 3.3, it shows that a L2 participant can make a significant distinction on an acoustic variable between NT2 and NT3 only if he/she made a distinction on the same variable between target-like T2 and T3, but not vice versa. No participant made a statistically reliable distinction on one measure between NT2 and NT3, but not between overt T2 and overt T3. This finding holds true for both of the elicitation.

3.2.4 Longitudinal Analysis

The production data were elicited at two different time points in an identical way in order to investigate whether any participants have made any progress in producing T2 and T3 between the two elicitations. This section discusses the finding of the two elicitations together by comparing the results of the six L2 speakers who participated in both of the elicitations.

Native Transcription. First, we compare the percentage of target-like transcription of the elicitations. Figure 3.3 presents the target-like productions, T2 and T3 altogether, scored base on the scoring criterion of Analysis 1, in which a target was scored as target-like when both of the transcriber perceived it as target-like, in both elicitations. The results are presented in bars of two different patterns for the two elicitations. Among the six participants, on one hand, five exhibited an increase in the target-like T2 and T3 production over the two elicitations. On the other hand, Participant WYZ showed an overall decrease in terms of the target-like targets transcribed by the transcribers. McNemar's test was conducted to determine the significance of the change

in the target-like percentage. The increase in the target-like production by FZT is statistically significant ($X^2(1, N=78)=5.0256, p<.05$), as well as those MXS ($X^2(1, N=78)=10.562), p<.05$) and LKL($X^2(1, N=78)=15.559, p<.05$). The decrease found in WYZ's target-like production is also significant ($X^2(1, N=78)=4.558, p<.05$).

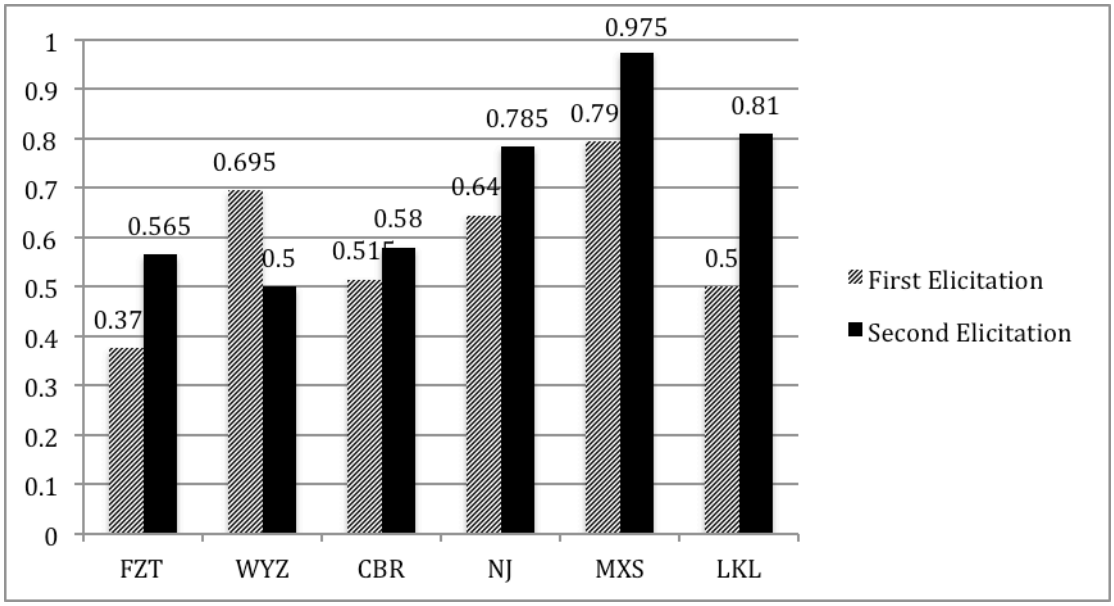


Figure 3.3 Percentage of target-like productions in Analysis 1

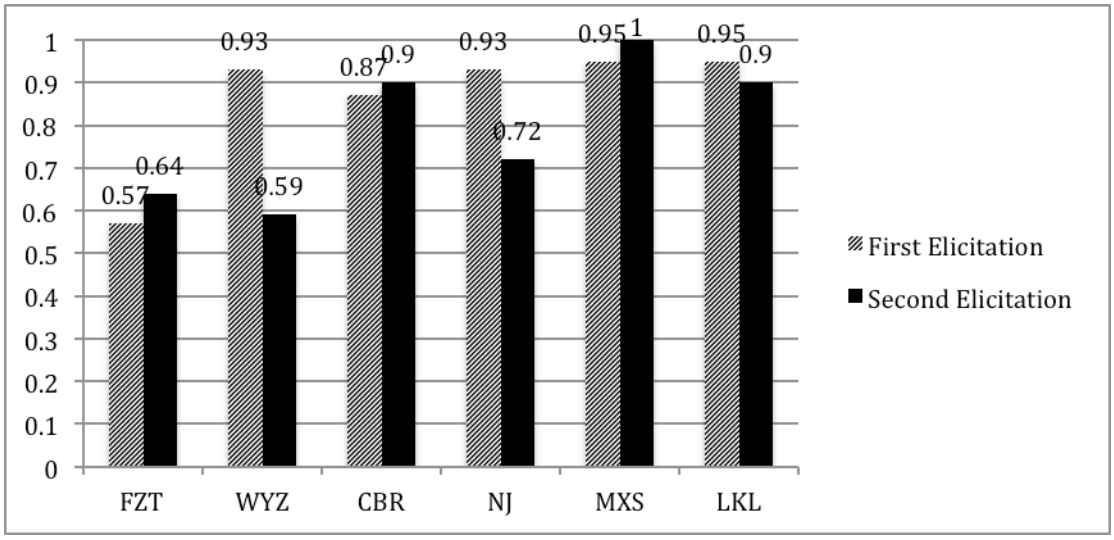


Figure 3.4 Percentage of target-like T2 in Analysis 1

Figure 3.4 illustrates the target-like T2 transcription scored base on the scoring criterion of Analysis 1. Since from the discussion earlier in the chapter, it was observed

that most participants produced T2 targets at a mostly target-like percentage, and thus have a relatively limited room for improvement, we assume little and limited change in the percentage of target-like T2 production. As seen from Figure 3.4, three participants (FTZ, CBR and MXS) exhibited a slight increase over time in the target-like T2 transcribed by the native transcribers, but McNemar's test shows that the changes are not significant. Three other participants (WYZ, NJ and LKL) have exhibited a decrease from the first elicitation to the second elicitation. While McNemar's test does not suggest that there is a significant decrease in LKL's target-like T2 production over time, it indicates the decrease found in NJ and WYZ's T2 production is statistically reliable. For WYZ, the target-like T2 percentage has decreased 34%, which is found significant ($X^2(1, N=39)=8.47, p<.05$). For NJ, the 21% decrease is also statistically reliable ($X^2(1, N=39)=4.08, p<.05$).

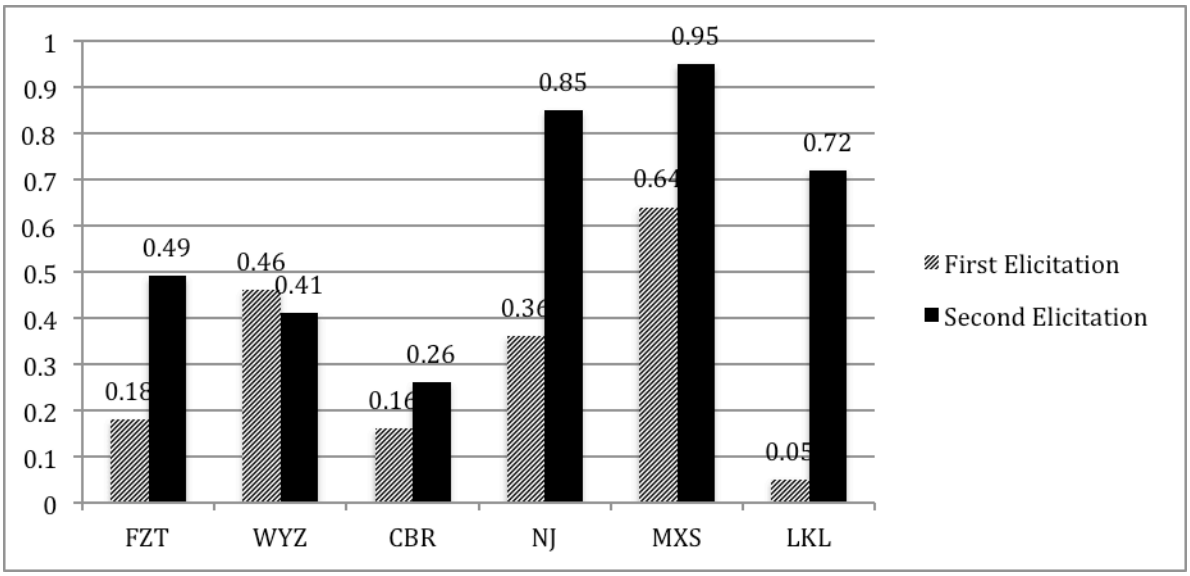


Figure 3.5 Percentage of target-like T3 in Analysis 1.

Figure 3.5 displays the results of target-like T3 across the elicitations in the same format as the previous two figures. Similar to what has been found from the longitudinal

results of T2 and T3 altogether, as shown in Figure 3.3, five participants exhibited an increase in the percentage of target-like T3 from the first to the second elicitation. Among the five of them, the increase observed in the production by FZT, NJ, MXS and LKL is statistically reliable. The increase CBR exhibited and the decrease found in WYZ's target-like T3 production have not been found significant. The detailed statistics of the McNemar's test conducted to compare the target-like production based on the strict transcribing criterion is reported in Table 3.4 below. Cells are shaded to indicate a statistically significant change from the first elicitation to the second one.

Table 3.4 The results of McNemar's test that compares the target-like production based on the strict scoring transcribing criterion

Participant	All	T2	T3
FZT	*X^2(1, N=78)=5.0256, p<.05*	*X^2(1, N=39)=0.19, p>.05*	*X^2(1, N=39)=6.72, p<.05*
WYZ	*X^2(1, N=78)=4.558, p<.05*).	*X^2(1, N=39)=8.47, p<.05*	*X^2(1, N=39)=0.03, p>.05*
CBR	*X^2(1, N=78)=0.7619, p>.05*).	*X^2(1, N=39)=0, p>.05*	*X^2(1, N=39)=0.75, p>.05*
NJ	*X^2(1, N=78)=2.702, p>.05*).	*X^2(1, N=39)=4.08, p<.05*	*X^2(1, N=39)=12.96, p<.05*
MXS	*X^2(1, N=78)=10.562, p<.05*).	N/A	*X^2(1, N=39)=8.64, p<.05*
LKL	*X^2(1, N=78)=15.559, p<.05*	*X^2(1, N=39)=0.16, p>.05*	*X^2(1, N=39)=22.32, p<.05*

Statistic Results on Acoustic Variables. Next, let us investigate the acoustic data of NT2 and NT3 statistically across the two elicitations. I used multiple linear regression models with two predictor variables to examine the interaction between the tones and the elicitations. Table 3.5 shows an example of the report of the main effects of tones and of elicitations, and the interaction effect between the tones and the elicitations of Participant FZT on the three acoustic variables. Each row reports the results of one acoustic variable. Within each row, the last column reports the

significance value of the effect of the row. For example, it was found that the tone effect is significant on F0 difference with a p-value of 0.0246, but there was no main effect of elicitation (*p=0.8943*) and no interaction between the segment and the elicitation (*p=0.355*). Essentially, in the longitudinal analysis, we are mainly concerned about the interaction of the tones and the elicitations because it informs us about whether the difference between NT2 and NT3 changes significantly from Time 1 to Time 2. From Table 3.5, regarding the acoustic variable of Duration, there was a main effect of tone and elicitation, but no interaction between the tone and the elicitation. It suggests that there was a significant distinction found between NT2 and NT3 and between the first and the elicitation on the duration, but the effect of tone is independent of the elicitations. On offset pitch, no main effect of tone was found, but there was a main effect of elicitation and an interaction between the tone and the elicitation, indicating that the distinction between NT2 and NT3 changed significantly from the first elicitation to the second elicitation.

Among the six participants, only one participant has shown an interaction between the tone and the elicitation on one acoustic variable, and that is the participant shown in Table 3.5, FZT. She exhibited a significant change of the difference between NT2 and NT3 on offset pitch from the first elicitation to the second one. Figure 3.6 illustrates this interaction. In the first elicitation, the difference between NT2 and NT3 is not significant, but the difference increased significantly in Time 2 as the two lines in the figure is more farther apart from each other.

Table 3.5: The results of the linear model constructed for the three acoustic variables from Participant FZT's longitudinal productions.

Difference		Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	40.059	9.002	4.450	2.78e-05 ***
	SegmentTone3	25.524	11.139	2.291	0.0246 *
	ElicitationTime2	1.786	13.395	0.133	0.8943
	SegmentTone3:ElicitationTime2	15.866	17.069	0.930	0.355
Duration		Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	0.29872	0.03473	8.601	5.85e-13 ***
	SegmentTone3	0.09379	0.04298	2.182	0.032053 *
	ElicitationTime2	0.18418	0.05168	3.564	0.000624 ***
	SegmentTone3:ElicitationTime2	-0.10208	0.06586	-1.550	0.125120
Offset		Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	303.630	8.528	35.603	<2e-16 ***
	SegmentTone3	-6.416	10.553	-0.608	0.5450
	ElicitationTime2	-27.788	12.690	-2.190	0.0315 *
	SegmentTone3:ElicitationTime2	-39.269	16.171	-2.428	0.0174 *
	Signif. codes: 0 '***' 0.01 '*'				

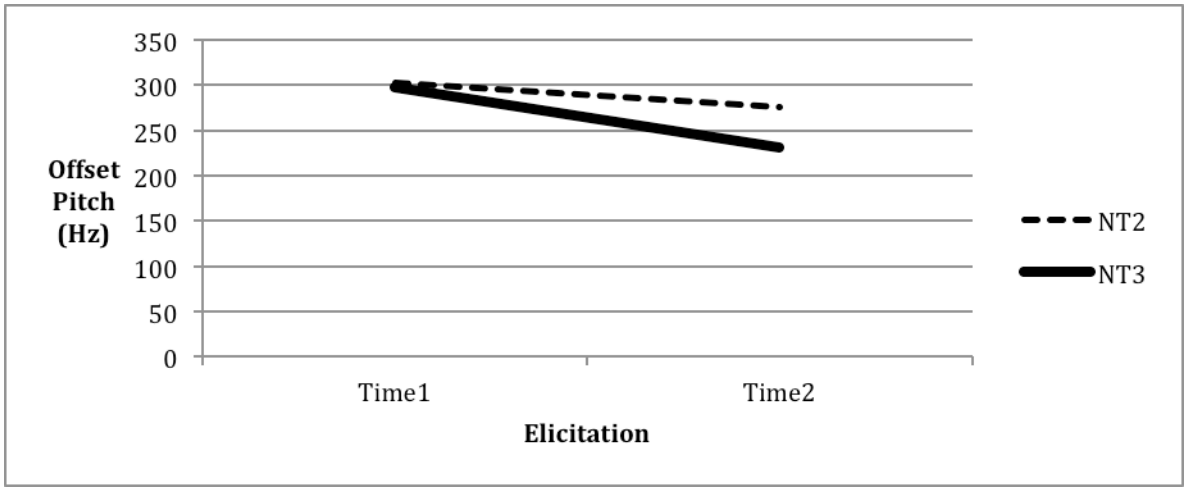


Figure 3.6 The interaction of tones and elicitation on offset pitch of Participant FZT

No other participants have demonstrated such an interaction between the tone and the time on any of the acoustic variables. It indicates that for the other participants, the distinction between NT2 and NT3 did not change significantly from Time 1 to Time 2. The detailed results of the linear models constructed for each individual were reported in Appendix C.

3.2.5 Interim Summary

In summary, Analysis 1 adopts a stricter method in scoring the transcription, in which a target is considered target-like when both of the native transcribers perceived it as target-like. Results from both of the elicitations show that most participants produced more target-like T2 than target-like T3 while two participants from the second elicitation produced a higher percentage of target-like T3 than target-like T2. In addition, from the linear model conducted for each individual, four of the ten participants in the first elicitation produced targets that were not transcribed as target-like but a significant distinction was made between NT2 and NT3 on at least one of the acoustic variables. In the second elicitation, three participants made a statistically reliable distinction between the non-target-like tones on at least one of the measures. This indicates that while some participants produced tones that were not perceived overtly by the transcribers, their non-target-like productions actually show some acoustic distinction to a certain degree that is statistically significant. Also, it is observed that all participants in both elicitations produced target-like contrast, which always demonstrates a statistically reliable distinction in F0 difference, but not necessarily the other two acoustic variables.

Next, from the longitudinal analysis on the native transcription, it was found that three of the six participants have made a statistically reliable improvement on producing target-like T2 and T3, combined. Specifically, for them, the improvement was from the statistically more target-like T3 produced in the second elicitation than in the first one. Interestingly, two participants have produced significantly less target-like T2 at Time 2, as compared to Time 1. Lastly, one participant has exhibited an interaction between the

tone and the time on one of the three acoustic variables, the offset pitch, suggesting that the distinction between NT2 and NT3 produced by this participant has changed significantly from Time 1 to Time 2 on offset pitch.

3.3 Analysis 2

3.3.1 Analysis Method

As stated previously, Analysis 2 differs from Analysis 1 in that a criterial threshold was established for L2 learners' target-like production. In order to decide on the basis of the transcription whether a L2 learner has systematically produced the contrast between T2 and T3, a participant's score on the transcription has to reach the 80% criterial threshold of being target-like for both T2 and T3. That is, a L2 participant was considered to have the contrast between T2 and T3 only when at least 80% of the targets were perceived by the native speakers. For L2 participants whose target-like production fails to reach the 80% criterion on one or both of the tones, their interlanguage (IL) is judged as lacking the contrast. In this case, all intended T2 and T3 productions were included in the statistical analysis, regardless of whether the tokens were perceived as target-like or not.

In addition to setting the threshold of 80% criterion of target-like transcription, the criterion for scoring a token as target-like in Analysis 2 is different from Analysis 1. In Analysis 1, the criterion for scoring a token as target-like was originally that at least one of the native transcribers heard it as intended. But Analysis 1 did not adopt this criterion because of the insufficient data points for statistical tests under this criterion. In Analysis 2, the issue of unequal sample sizes, as encountered in Analysis 1, was not

present since all intended T2 and T3 were included in the statistical test when a participant's target-like production failed to reach the 80% criterion.

The last methodological difference between Analysis 1 and 2 is the statistical methods. As mentioned in Analysis 1, paired-sample t-tests were not used in Analysis 1 because it cannot analyze data with unequal sample sizes without compromising some data points. However, Analysis 2 takes all data produced by a L2 participant into account, so there is no concern of unequal sample sizes. As a result, I used paired-sample t-test in Analysis 2, in deciding whether a participant made a statistically reliable distinction on the acoustic variables between T2 and T3. Also, if a statistically reliable distinction can be shown and by two different statistical methods, the occurrence of covert contrast could be better supported. All of the other factors regarding the method of Analysis 2 remain identical to the ones in Analysis 1.

3.3.2 Results: First Elicitation

Native Transcription. Analysis 2 adopts a looser scoring criterion than Analysis 1 for the native transcription, so it is expected that the target-like percentage will be higher than Analysis 1. Figure 3.7 below shows the percentage of targets that were transcribed as target-like according to the looser scoring criterion. All of the gray striped bars are reaching near the top, indicating that most T2 targets produced by the participants were transcribed as target-like. The T3 results, plotted by the solid bars, are showing more participants reaching a higher percentage of target-like transcription. Recall that from Analysis 1, only one participant (MXS) produced more 50% target-like

T3. By scoring the transcription in a looser criterion in Analysis 2, Figure 3.7 shows that five participants (FZT, CBR, WYZ, MXS and NJ) produced more than 50% of T3 that were transcribed as target-like, and another one participant (XLX) has near 50% of target-like T3 production. However, what remains unchanged is that the target-like percentage of T2 is still higher than that of T3. It indicates that the regardless of how the transcription was scored, T2 were transcribed as target-like more than T3, thus suggesting it is more difficult for L2 learners to produce target-like T3 than target-like T2.

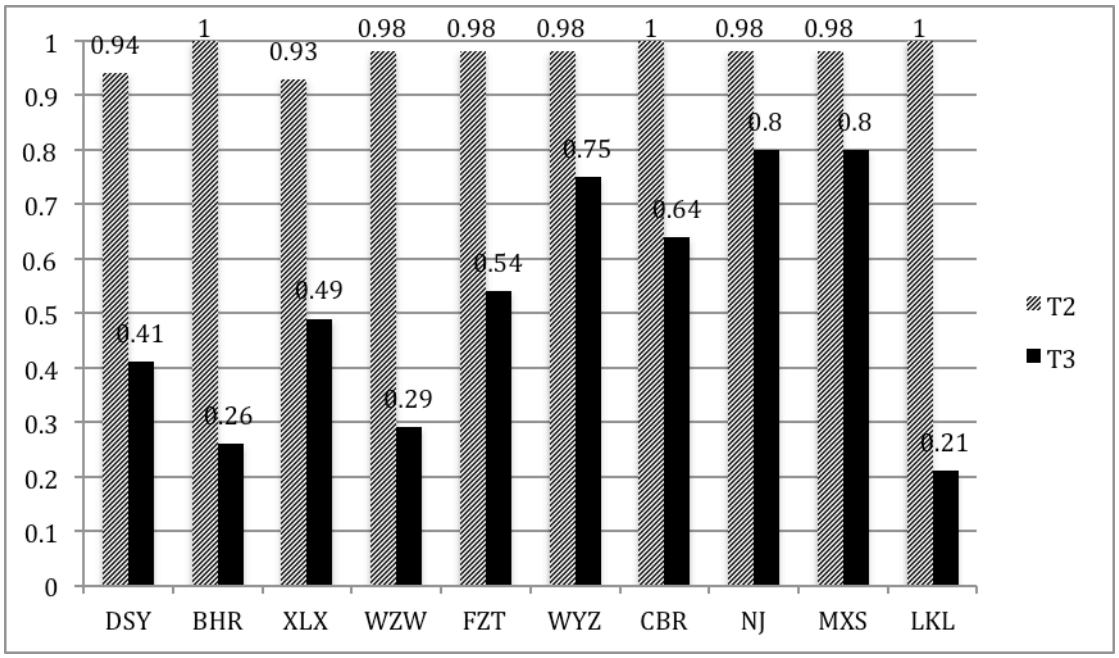


Figure 3.7 Percentage of targets that were transcribed as target-like by the looser scoring criterion (First elicitation: November 2013)

With the 80% criterial threshold of target-like production transcribed by the native judges, two participants (MXS and NJ) have been crediting with having acquired the contrast. Both of their production of T2 and T3 have reached at least 80% of being perceived as target-like. Their production has exhibited overt contrast between T2 and T3. The other eight participants' target-like T3 percentage did not reach 80%. Thus, they

were not considered as having acquired the contrast and the purpose of the statistical analysis was to determine whether the participants lacking overt contrast made a statistically reliable distinction between the targets on the acoustic variables.

Statistical Results. Table 3.6 below presents the results of the pair-sample t-tests that compare all T2 and T3 on each acoustic variable for each participant. The first column lists the participants. The data of the two participants with overt contrast are presented together at the bottom of the table and the rest of the participants' are shown above them. Each of the rest of the columns reports the t-test result on an acoustic variable for the participant of the row. The cells are shadowed to indicate a statistically reliable distinction between all T2 and T3 on the specific acoustic variable. To exemplify the numbers in the table, we look at the first row, results of Participant CBR. The distinction between the pitch difference of all T2 and all T3 is significant with a p value of 0.035. On both the normalized duration and offset pitch height, the distinction between the T2 and T3 he produced is also statistically reliable with a significance value of 0.

The t-test models we constructed show that more participants have made a statistically reliable distinction between all T2 and T3 on the duration variable than on the other two. There were nine participants making a statistically reliable distinction on the normalized duration, consisting of both ones who exhibited an overt contrast and all but one of the participants whose production lacks of an over contrast. Six of the nine participants who made a statistically reliable distinction on the duration also implemented a significant distinction on the offset pitch. In terms of the F0 difference, six participants, four without an overt contrast and both ones exhibiting an overt contrast, have made a statistically reliable distinction.

Table 3.6: Paired sample t-test results and significant value of all T2 and all T3 for each L2 speaker			
Participant	**F0 difference**	**Normalized Duration**	**Offset Pitch Height**
Participants without overt contrast			
CBR	*t(38)=-2.186, p=.035, p<.05*	*t(38)=-7.478, p=0, p<.05*	*t(38)=4.951, p=0, p<.05*
DSY	*t(38)=-1.942, p=.06, p>.05*	*t(38)=-5.539, p=0, p<.05*	*t(38)=5.523, p=0, p<.05*
BHR	*t(38)=-1.223, p=.229, p>.05*	*t(38)=-4.126, p=0, p<.05*	*t(38)=4.268, p=0, p<.05*
XLX	*t(38)=-1.948, p=.059, p>.05*	*t(38)=-3.393, p=.002, p<.05*	*t(38)=-.901, p=.373, p>.05*
FZT	*t(38)=-3.601, p=.001, p<.05*	*t(38)=-3.469, p=.001, p<.05*	*t(38)=4.005, p=0, p<.05*
LKL	*t(38)=-2.15, p=.038, p<.05*	*t(38)=-2.942, p=.006, p<.05*	*t(38)=1.862, p=.07, p>.05*
WYZ	*t(38)=-4.532, p=0, p<.05*	*t(38)=-1.698, p=.098, p>.05*	*t(38)=1.57, p=.125, p>.05*
WZW	*t(38)=-1.763, p=.086, p>.05*	*t(38)=-3.422, p=.002, p<.05*	*t(38)=3.079, p=.004, p<.05*
Participants exhibiting overt contrast			
NJ	*t(38)=-7.112, p=0, p<.05*	*t(38)=-3.302, p=.002, p<.05*	*t(38)=1.429, p=.161, p>.05*
MXS	*t(38)=-5.8, p=0, p<.05*	*t(38)=-6.604, p=0, p<.05*	*t(38)=7.04, p=0, p<.05*

Table 3.7 summarized the results shown in Table 3.6 with the percentage of target-like productions (as shown in Figure 3.7), arranged according to whether the L2 participants produced an overt contrast, based on the 80% threshold of the transcriptions, and whether acoustic analysis show a statistically reliable distinction. The acoustic variables are arranged in the order of the hypothesized trajectory of implementing the contrast, which states that a L2 participant will make a statistically reliable distinction on the duration before making one on the offset pitch and the F0 and that a L2 participant will implement the contrast covertly on the offset pitch before making a reliable distinction on the F0 difference. Therefore, on the right-hand side of the table, the acoustic variables are listed from left to right as duration, offset pitch, and lastly, F0 difference. In addition, for the ease of data presentation and interpretation, the participants are rearranged based one how they implemented the contrast acoustically

according to the hypothesized trajectory. That is, Table 3.7 lists the participants in the order of whether their acoustic production follows the trajectory. Participants whose production generally follows the proposed order are listed before the others from the bottom of the table. The two participants exhibiting an overt contrast are presented at the bottom.

	Table 3.7: L2 speakers' production, according to transcriptions and acoustic measures					
	Percentage of target-like production (80% threshold)			**Acoustic measures with statistical results**		
	T2	**T3**	**Transcription**	**Duration**	**Offset pitch**	**F0 Diff.**
Participants without overt contrast						
WYZ	98%	75%	No	-	-	+
XLX	93%	49%	No	+	-	-
LKL	100%	21%	No	+	-	+
WZW	98%	29%	No	+	+	-
DSY	94%	41%	No	+	+	-
BHR	100%	26%	No	+	+	-
FZT	100%	54%	No	+	+	+
CBR	98%	64%	No	+	+	+
Participants exhibiting overt contrast						
NJ	98%	80%	Overt	+	-	+
MXS	98%	80%	Overt	+	+	+

To exemplify how to read Table 3.7, we look at Participant WYZ and CBR. WYZ did not produce an overt contrast between T2 and T3 because his target-like T3 production does not reach 80% while 98% of his T2 were perceived as target-like. He is listed on the top above all the other participants because he produced a statistically

reliable distinction between T2 and T3 on the measure of F0 difference, but not on the other two measures, which was not predicted by the proposed trajectory. On the other hand, CBR's result is listed at the bottom along with the participants without an overt contrast because his T2 and T3 measurements are statistically different from each other on all three acoustic variables, which follow the hypothesized trajectory.

Now, let us look at the results in detail. From Table 3.7, it is found that all L2 speakers made a contrast covertly in at least one of the three acoustics measures. Three participants (MXS, CBR, and FZT) produced a statistically reliable distinction on all three variables. Among them, MXS's productions were systematically transcribed as target-like by the native transcribers. Five participants exhibited a covert contrast in two acoustic measures. Three of them (BHR, DSY and WZW) demonstrated a reliable distinction in both the duration measure and offset pitch. The other two (NJ and LKL) show the contrast in the F0 difference and the normalized duration variables. Surprisingly, NJ's production was overtly perceived as target-like by the native transcribers while a statistically reliable distinction was found in two acoustic variables, instead of three like MXS's. It might imply a suggestion similar to what Analysis 1 finds, that a L2 participant who systematically exhibits an overt contrast between T2 and T3 does not necessarily have to implement a statistically reliable distinction on all three acoustic variables. Lastly, two participants show the distinction only on one acoustic measure; XLX contrasted T2 and T3 in the normalized duration while WYZ made a significant distinction in the F0 drop occurring before the turning point.

Regarding the order of acoustic variables in which L2 participants made a significant distinction, all six participants (MXS, CBR, FZT, BHR, DSY and WZW) who

made a statistically reliable separation on the offset pitch also showed a significant distinction on the normalized duration to the turning point. However, not all participants who implemented the contrast on the F0 difference also showed a similar distinction on the other two variables. Among the six who implemented the contrast on the F0 difference, half of them (MXS, CBR and FZT) also have made a reliable distinction on the first two acoustic variables. The other three did not significantly implement the contrast on the offset pitch, including a participant, WYZ, whose production neither show any statistic difference between T2 and T3 on the durational variable. In other words, three out of ten L2 participants did not produce the targets that precisely follow the trajectory I proposed but the other seven did.

To sum up, in the first elicitation, Analysis 2 indicates that two participants exhibited a covert contrast based on the 80% threshold for acquisition. The other eight participants all made a statistically significant distinction on at least one of the acoustic measures. The next section presents the same analysis of the second elicitation where six participants from the first elicitation also participated, with another six participants.

3.3.3 Results: Second Elicitation

Native transcription. Figure 3.8 on the next page shows results by scoring the native transcriptions in the looser method. Similarly to the first elicitation, the results were shown in percentage of the T2 and T3 targets that were transcribed as target-like. A similar pattern was found from the previous elicitation; in general, most participants' T2 targets reach near target-like transcription. All of the participants produced T2 that were transcribed as target-like more than 90% of the attempts. Three participants' T2

(JK, CBR and MXS) were all perceived as target-like. Also, it is shown that the ten of the twelve participants have a higher percentage of target-like T2 than target-like T3; two (LJ and NJ) produced more target-like T3 than T3. This result agrees with what has been found in Analysis 1. The different scoring criterion does not change this finding, that most L2 learners produced target-like T2 more easily than target-like T3.

Observing the percentage individually, some participants' transcriptions show similar results. First, the results of KL, JK and FMY seem to behave similarly. While their T2 were transcribed as target-like near or over 95% of the time, the T3 they produced were only perceived as target-like by chance, even with the criterion that a target is considered target-like when at least one of the two transcribers perceived it as target-like. HL and LWQ's productions show similar transcription results. Both of their T2 targets attain nearly 95% of target-like attempts, and about 70% of the T3 were overtly perceived. Since the five participants discussed above failed to produce target-like T3 at a minimum of 80% of the attempt, it is concluded that they have not acquired the T2-T3 contrast.

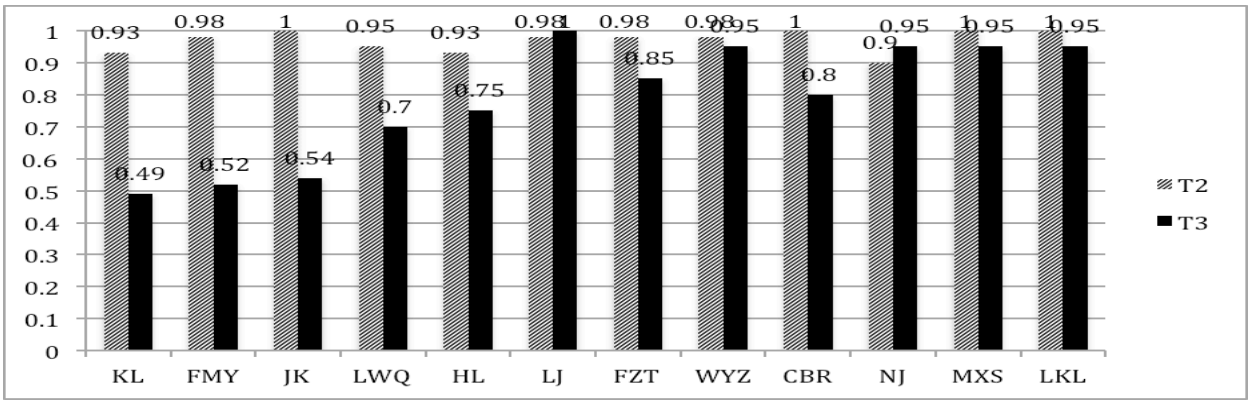


Figure 3.8 Percentage of targets that were transcribed as target-like by the looser scoring criterion. (Second elicitation: April 2015)

The other seven participants were all credited with having acquired the T2 and T3 contrast with the 80% criterial threshold of target-like production transcribed by the native judges. All of their productions of T2 and T3 have reached at least 80% of being perceived as target-like; therefore, we consider that their production has exhibited overt contrast between T2 and T3. Below I present the statistical analysis investigate whether the participants lacking an overt contrast made a statistically reliable distinction on the acoustic variables and how the statistically reliable distinctions on acoustic variables were implemented by each participant.

Statistical Analysis. Table 3.8 reports the paired-sample t-test results with the significance value p, for each participant on each acoustic variable. The participants were grouped and listed on the basis of whether they have exhibited an overt contrast according to the 80% criterial threshold. The seven participants whose production exhibits an overt contrast between T2 and T3 are listed in the bottom of the table whereas the data of the five who lack of the contrast are presented on the top. With in the groups, with or without an overt contrast, the participants are organized in a sequence determined by whether the production follows the trajectory of implementing the distinction on the acoustic variables. Cells of the table were shaded when the p-value indicates a significant result.

Table 3.9 combines Table 3.8 with the percentage of productions that were transcribed as target-like for each individual participant. First, we look at the participants exhibiting an overt contrast. Among seven of them, four participants (CBR, NJ, MXS and LJ) have made a statistically reliable distinction between T2 and T3 on all three

acoustic variables. One of them (CBR) produced target-like T3 at 80% of the attempts while the other three participants' T3 productions were overtly perceived more than 95% of the time. The other three participants (LKL, FZT and WYZ) with an overt contrast have implemented a significant distinction between T2 and T3 on two of the three acoustic variables. Specifically, LKL and FZT made a statistically reliable distinction on the offset pitch and the F0 difference. WYZ's T2 and T3 productions contrasted each other significantly on the duration variable and the F0 difference. What all seven participants credited with an overt contrast share is that all the T2 and T3 they produced contrast with each other on the F0 difference. It corresponds with what was found in Analysis 1, that all the overt T2 and T3 statistically predict the F0 difference produced by each participant.

Table 3.8: Paired sample t-test results and significant values of all T2 and all T3 for each L2 speaker in the second elicitation

Participant	F0 difference	Normalized Duration	Offset Pitch Height
Participants without overt contrast			
KL	*t(38)=-1.28, p=0.2, p>.05*	*t(38)=0.04, p=0.9, p>.05*	*t(38)=-0.6, p=0.5, p>.05*
HL	*t(38)=-2.75, p=0.008, p<.05*	*t(38)=0.2, p=0.7, p>.05*	*t(38)=0.01, p=0.9, p>.05*
JK	*t(38)=-0.63, p=0.5, p>.05*	*t(38)=-3.43, p=0.006, p<.05*	*t(38)=0.19, p=0.84, p>.05*
LWQ	*t(38)=-6.98, p=0, p<.05*	*t(38)=-5.68, p=0, p<.05*	*t(38)=0.8, p=.38, p>.05*
FMY	*t(38)=-4.11, p=0, p<.05*	*t(38)=-2.83, p=.007, p<.05*	*t(38)=1.28, p=0.2, p>.05*
Participants exhibiting overt contrast			
LKL	*t(38)=-5.79, p=0, p<.05*	*t(38)=-1.5633, p=0.12, p>.05*	*t(38)=6.13, p=0, p<.05*
FZT	*t(38)=-4.32, p=0, p<.05*	*t(38)=-1.95, p=0.058, p>.05*	*t(38)=3.71, p=0, p<.05*
WYZ	*t(38)=-3.32, p=0.001, p<.05*	*t(38)=-3.81, p=0, p<.05*	*t(38)=1.02, p=0.3, p>.05*
CBR	*t(38)=-5.55, p=0, p<.05*	*t(38)=-5.17, p=0, p<.05*	*t(38)=2.32, p=0.02, p<.05*
NJ	*t(38)=-6.657, p=0, p<.05*	*t(38)=-2.45, p=0.01, p<.05*	*t(38)=-4, p=0, p<.05*
MXS	*t(38)=-9.09, p=0, p<.05*	*t(38)=-2.612, p=0.01, p<.05*	*t(38)=3.03, p=0.004, p<.05*
LJ	*t(38)=-8.5, p=0, p<.05*	*t(38)=-8.37, p=0, p<.05*	*t(38)=7.28, p=0, p<.05*

Table 3.9: L2 speakers' production, according to transcriptions and acoustic measures in the second elicitation						
	Percentage of target-like production (80% threshold)			Acoustic measures with statistical results		
	T2	T3	Transcription	Duration	Offset pitch	F0 Diff.
Participants without an overt contrast						
KL	93%	49%	No	-	-	-
HL	93%	75%	No	-	-	+
JK	100%	54%	No	+	-	-
LWQ	95%	70%	No	+	-	+
FMY	98%	52%	No	+	-	+
Participants exhibiting an overt contrast						
LKL	100%	95%	Overt	-	+	+
FZT	98%	85%	Overt	-	+	+
WYZ	98%	95%	Overt	+	-	+
CBR	100%	80%	Overt	+	+	+
NJ	90%	95%	Overt	+	+	+
MXS	100%	95%	Overt	+	+	+
LJ	98%	100%	Overt	+	+	+

Now, we proceed to the statistical results of the five participants whose productions fail to show an overt contrast between T2 and T3. On one hand, four of them made a statistically reliable distinction on at least one of the acoustic variables. Particularly, LWQ and FMY made a significant difference on two of them, the normalized duration and the F0 difference. JK and HL statistically implemented the contrast on one yet different acoustic variable; JK's T2 and T3 productions contrast in

duration whereas HL made statistically reliable distinction on the F0 difference. On the other hand, there is one participant, KL, whose productions do not exhibit any statistical distinction on neither one of the three acoustic variables investigated in the current study. With her production showing no statistically significant contrast, she has not been credited with making the T2-T3 contrast covertly.

In addition, from the five participants who did not exhibit an overt contrast, no one has made a reliable distinction on all three acoustic variables, nor has one of them implemented the contrast acoustically on the offset pitch variable. One thing notable to be mentioned is that HL and LWQ's T3 target-like production is nearly the 80% criterial threshold, respectively 75% and 70%, they also both implemented a reliable contrast on the F0 difference, which is found to be the acoustic variable implemented distinctively by all participants with an overt contrast.

3.3.4 Longitudinal Analysis

Similar to Analysis 1, this section compares and reports the results of the first and the second elicitation in Analysis 2. We conducted McNemar's test to see if the native transcriptions have changed significantly from Time 1 to Time 2. In addition, since we have set the 80% criterial threshold for target-like transcription to determine whether a participant has acquired the contrast or not, we were able to investigate whether a L2 participant has made a progress from having no contrast between T2 and T3 to implementing the target tones covertly, or from implementing a covert contrast to being credited with acquiring the contrast. Lastly, we compare the acoustic variables in which

each participant has made a statistically reliable difference in Time 1 with those in Time 2. In this case, we can understand how a participant's acoustic implementation of the contrast differs across the elicitations as he/she made a progress from having no contrast to making a covert contrast or from making a covert contrast to implementing the contrast overtly.

Native Transcription. We now look at the longitudinal change on the target-like productions scored based on the loose scoring criterion, which a target is considered target-like when at least one transcriber perceived it as target-like. Figure 3.9, 3.10 and 3.11 show the percentage of target-like productions, respectively of all targets, for T2, and for T3, categorized by the elicitations. Table 3.10 reports the detailed results of the McNemar's test that were conducted to determine whether the change, if there is any, in the target-like production is significant or not.

From Figure 3.9, when we look at the percentage of target-like T2 and T3 combined, all six participants seemed to have made a progress in that all of their percentage of target-like production has increased. Each of them has produced more targets that were perceived as target-like by the transcribers. Two of the participants produced a statistically reliable increase from the first elicitation to the second elicitation. LKL produced 31% more target-like T2 and T3 combined, which is a significant increase ($X^2(1, N=78)=27.034, p<.05$). The McNemar's test also shows significance for MXS's 3.5% increase ($X^2(1, N=78)=4, p<.05$). No significance was found for other participants' results.

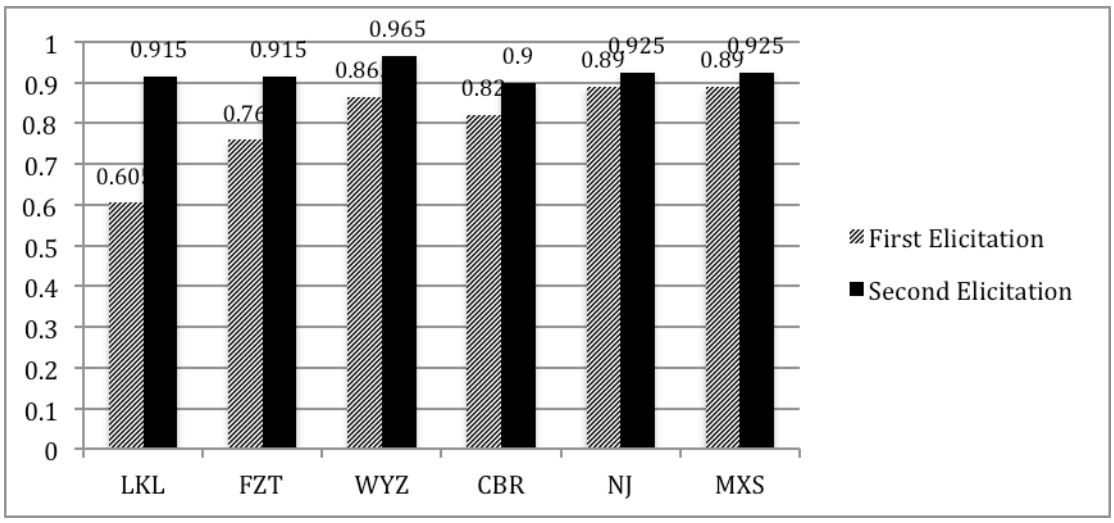


Figure 3.9 Comparison of percentages of target-like productions of T2 and T3 together scored based on the loose criterion

Table 3.10. The results of McNemar's test that compares the target-like production based on the loose scoring transcribing criterion.			
Participant	All	T2	T3
FZT	$X^2(1, N=78)=3.3684, p>.05$	*N/A*	$X^2(1, N=39)=3.76, p=.052, p>.05$
WYZ	$X^2(1, N=78)=3.5, p>.05$).	$X^2(1, N=39)=0, p>.05$	$X^2(1, N=39)=4.08, p<.05$
CBR	$X^2(1, N=78)=1.25, p>.05$).	*N/A*	$X^2(1, N=39)=1.25, p>.05$
NJ	$X^2(1, N=78)=0.307, p>.05$).	$X^2(1, N=39)=0.8, p>.05$	$X^2(1, N=39)=3.125, p>.05$
MXS	$X^2(1, N=78)=4, p<.05$).	N/A	$X^2(1, N=39)=3.125, p>.05$
LKL	$X^2(1, N=78)=27.034, p<.05$	*N/A*	$X^2(1, N=39)=27.03, p<.05$

Figure 3.10 and 3.11 show the comparison of target-like T2 and T3 for each participant in the two elicitations. As seen from Figure 3.10, there is no visibly significant decrease or increase in the percentage of target-like T2. Four of the six participants maintained the same target-like percentage of T2. As shown in the middle column of Table 3.9, the 8% decrease found in NJ's T2 production and the 2% increase of MXS's result is not statistically significant.

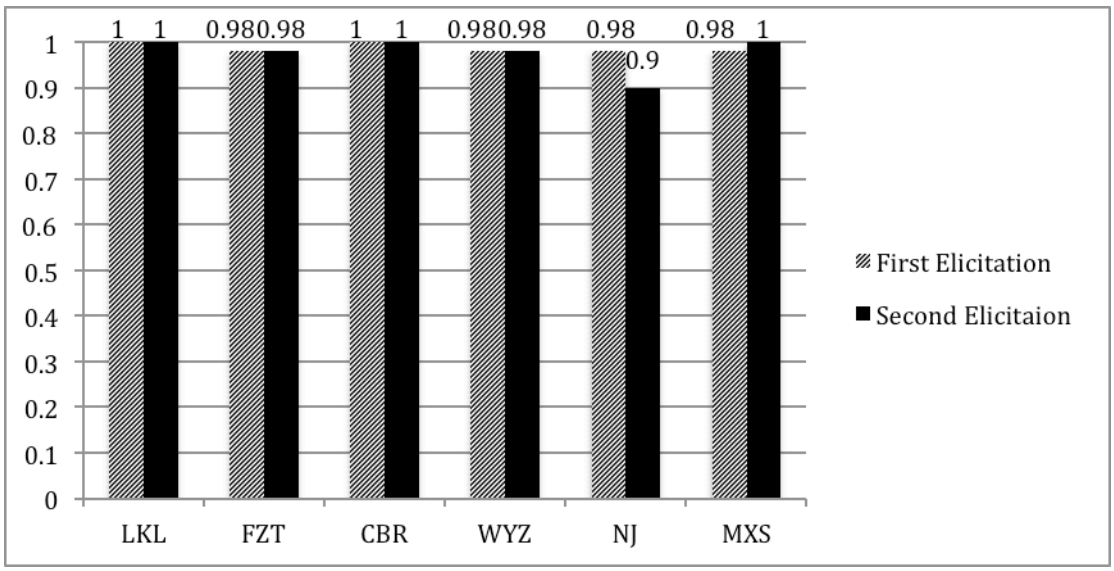


Figure 3.10 Comparison on the percentages of target-like T2 in Analysis 2.

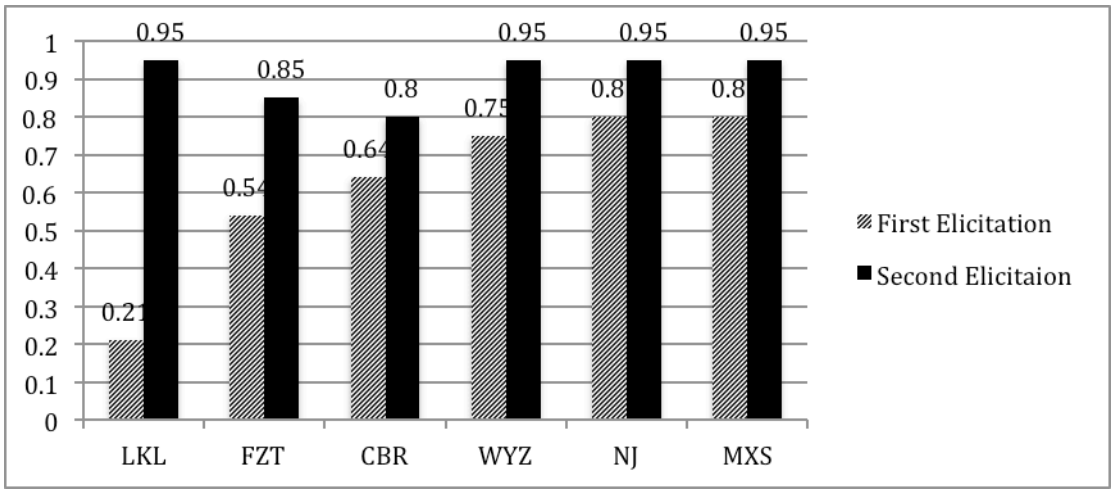


Figure 3.11 Comparison on the percentages of target-like T3 in Analysis 2.

From Figure 3.11, it is found that when the transcription was scored in the criterion of Analysis 2, all participants have exhibited an increase in target-like T3. Specifically, LKL has produced 74% more target-like T3 in the second elicitation than in the first one. The increase for her is found to be significant ($X^2(1, N=39)=27.03, p<.05$). WYZ's production has demonstrated a statistically significant improvement ($X^2(1, N=39)=4.08, p<.05$) in that 20% more T3 were scored as target-like in the second

elicitation. For FZT, she has shown a 31% increase of target-like T3 from the first to the second elicitation. The increase is, although not statistically significant, actually near significant (*p=0.52*), as seen from the last column in Table 3.10.

In conclusion, regarding the results of the native transcription, five of the six participants (FZT, LKL, MXS, NJ and WYZ) have produced significantly more target-like T3 in second elicitation than in the first one. Among them, four exhibited the increase when the transcription was scored in a way that a target is scored as target-like when both transcribers perceived it as target-like, and one shows the increase when the transcription was scored in the method that at least one transcriber had to perceived it as target-like. In terms of longitudinal change in producing T2, two participants (NJ and WYZ) produced significantly less target-like tokens when the transcription is scored by the strict criterion. The next section discusses what's observed from the longitudinal results in terms of whether the participants' target-like production reach 80% of the attempts, and thus if they have made any progress from implementing the T2-T3 contrast covertly to making an overt contrast.

Progress from Covert to Overt Contrast Below, Table 3.11 shows the six participants' results of the target-like production percentage from both elicitations. The left side of the table presents the results of the first elicitation with whether the participant produced an overt contrast based on the 80% threshold of the transcription. The right side mirrors the presentation of the left, showing the results of the second elicitation for the purpose of comparison. Cells are shaded to indicate an overt contrast produced systematically by a participant at the minimum of 80% of attempts.

Table 3.11: L2 speakers' production, according to transcriptions and acoustic measures						
	Percentage of target-like production (80% threshold)					
	First Elicitation			Second Elicitation		
	T2	T3	Transcription	T2	T3	Transcription
WYZ	98%	75%	No	98%	95%	Overt
LKL	100%	21%	No	100%	95%	Overt
FZT	100%	54%	No	98%	85%	Overt
CBR	98%	64%	No	100%	80%	Overt
NJ	98%	80%	Overt	90%	95%	Overt
MXS	98%	80%	Overt	100%	95%	Overt

It is shown that in the first elicitation, among the six participants, four participants listed on the top did not get credited with producing an overt contrast as their target-like T3 failed to reach the 80% threshold. Two (NJ and MXS) have exhibited an overt contrast between T2 and T3. However, it is found that from the data elicited seventeen months later, all of the six participants have produced target-like T3 at least 80% of the attempts while their target-like T2 production remain near to 100%. That is, while the two participants who have produced an overt contrast in the first elicitation still remain producing an overt contrast, the other four participants have made a progress from failing to produce an overt contrast to exhibiting an overt contrast between T2 and T3.

Acoustic Implementation of the Contrast. Table 3.12 summarizes the transcription results with the statistical results of the acoustic variables.

Table 3.12. Both Elicitation: L2 speakers' production, according to transcriptions and acoustic

measures							
	Elicitation	Percentage of target-like production (80% threshold)			Acoustic measures with statistical results		
		T2	T3	Transcription	Duration	Offset pitch	F0 Diff.
LKL	First	100%	21%	No	+	-	+
	Second	100%	95%	Overt	-	+	+
FZT	First	100%	54%	No	+	+	+
	Second	98%	85%	Overt	-	+	+
WYZ	First	98%	75%	No	-	-	+
	Second	98%	95%	Overt	+	-	+
CBR	First	98%	64%	No	+	+	+
	Second	100%	80%	Overt	+	+	+
NJ	First	98%	80%	Overt	+	-	+
	Second	90%	95%	Overt	+	+	+
MXS	First	98%	80%	Overt	+	+	+
	Second	100%	95%	Overt	+	+	+

The first column lists the six participants. There are two rows of results for each participant. With the results of the first elicitation shown in the first row of each participant, the results of the second elicitation are presented in the row directly below for the convenience of comparing. Overall, throughout the two elicitations, it is found that all the six participants have made a statistically reliable distinction between all the T2 and T3 targets on the F0 difference. Among them, two (CBR and MXS) of the participants' acoustic measurement shows no change in which acoustic variables present a reliable distinction. Both of them acoustically implemented the contrast on all

the three variables in the first and the second elicitation. The other four participants present different patterns of the acoustic implementation of the contrast, which I will discuss in the following paragraphs.

First, Participant LKL made a statistically reliable distinction between T2 and T3 on the duration to the turning point and F0 difference in the first elicitation, but it was found that her second production did not show such a distinction on the duration. Instead, a reliable difference between T2 and T3 was found on the offset pitch. As her production developed from implementing a covert contrast to exhibiting an overt contrast between T2 and T3, she made the contrast acoustically on two variables which always include F0 difference and another one, that changed across the elicitations. Secondly, FZT implemented the contrast on all three variables in the first elicitation, but did not show the similar reliable distinction on the duration variable in the production seventeen months later. Interestingly, her target-like T3 production has increased 31% (near significance, from 54% to 85%), but the number of the acoustic variable she implemented the contrast decrease from three to two. Next, Participant WYZ and NJ both made a statistically reliable distinction on one more acoustic variable in the second elicitation than in the first one. From WYZ's production, a significant distinction was found only on F0 difference in the first elicitation. In the second elicitation, he has also produced the reliable contrast on the duration to the turning point, in addition to F0 difference. Neither of his productions has shown any reliable difference between T2 and T3 on the offset pitch.

Lastly, we turn to NJ's results of the statistical analysis on the acoustic variables. NJ was one of the two participants who have exhibited an overt contrast in the first elicitation. In the first elicitation, she made a statistically reliable distinction on the duration and the F0 difference, but not on the offset pitch. In the second elicitation, when her production still exhibits an overt T2-T3 contrast systematically according to the 80% threshold, it was found that she has also acoustically implemented the contrast on the offset pitch, with a significant distinction on the other two variables that she initially have made. In other words, while the 15% increase of her target-like T3 was not considered statistically significant, she has made a progress that is statistically reliable from implementing the contrast on two acoustic variables to making a significant distinction on all three acoustic variables.

3.3.5 Interim Summary

Data gathered from the first elicitation shows that two participants have exhibited an overt contrast as both of their T2 and T3 productions were transcribed 80% or more of the attempts. The acoustic analysis of their productions showed that the T2 and T3 were statistically different from each other on two or all acoustic variables. The other eight participants were not credited with having acquired the T2-T3 contrast overtly because their target-like T3 did not reach the 80% criterial threshold. However, all of them made a statistically reliable distinction between the T2 and T3 attempts on at least one of the three acoustic variables measured in this study. That is, although the T2 and T3 they produced were not considered an overt contrast yet, they implemented the

contrast that is statistically significant but not systematically enough to be perceived as target-like by the native transcribed.

The second elicitation found seven participants exhibiting an overt contrast with the same 80% target-like criterion. All seven of them have made a statistically reliable distinction at the minimum of two acoustic variables, which always include a significant distinction on the F0 difference. Among the five participants whose productions do not exhibit an overt contrast between T2 and T3, four of them made a statistically reliable distinction on at least one acoustic variable. It is shown one participant shows no reliable acoustic distinction.

In terms of which acoustic variable was implemented covertly by most participants, the first elicitation showed that nine of the ten participants made a significant distinction on the normalized duration to the turning point, which is the most implemented variable in the first elicitation. The second elicitation found F0 difference to be the one that was most frequently implemented, with ten of the twelve participants making a reliable distinction.

Lastly, longitudinal analysis found that the four participants who did not show an overt contrast between T2 and T3 from the native transcription in the first elicitation exhibit an overt implementation of the contrast in the second elicitation. In the first elicitation, although they did not implement the T2-T3 contrast overtly, they made a statistically reliable distinction on one or more acoustic variables; thus, they could be considered to have exhibited a covert contrast. Therefore, these four participants have made a longitudinal progress from implementing the contrast covertly to exhibiting an

overt contrast. The two participants who have acquired an overt contrast in the first elicitation have remained at the overt stage in the acquisition of the T2-T3 contrast.

Chapter 4

Identification Task

This chapter describes the methodology of data collection and analysis, and the findings of the identification task. From the production test, as discussed in the previous chapter, the participants have shown different results in terms of whether they have produced an overt contrast or a covert contrast. Accordingly, the identification task, as well as the discrimination task in the next chapter, aimed at comparing the performance of the two groups of L2 participants, namely the overt group and the covert group. The identification task investigated whether such a difference, as found in their production, is also present in the participants' perception in identifying T2 and T3. For convenience, the hypothesis that the identification task aimed to test is repeated as the following:

(5) The L2 participants who produced an overt contrast between T2 and T3 will identify T2 and T3 statistically better than the L2 participants who implement the contrast covertly, who will then identify T2 and T3 statistically better than the participants who demonstrated no contrast at all.

The chapter will proceed as follow. Section 4.1 presents the detailed methodology of data collection and analysis, including participants' background, stimuli, procedures, the scoring method and the analysis of the data. Section 4.2 reports the results of the participant groups and statistical analysis of the comparison between the participant groups.

4.1 Methodology

4.1.1 Participants. The participants who were recruited for the second elicitation of the production task also participated in the perception task. There are twelve participants in total. From the results of the production task, the participants were grouped based on whether they implement the T2-T3 contrast overtly, or covertly, from Analysis 2, resulting in two groups of participants; one that shows an overt contrast (overt), and one that has yet to produce an overt contrast but has shown a statistically reliable distinction (covert). Among the twelve participants, seven participants were grouped into the overt participant group, and four belong to the covert group. There was one participant who produced no contrast at all, whose production I have excluded in the statistical test.

4.1.2 Stimuli. The targets that were used in the perception task were identical to those in the production experiment, which are 78 CV-structured monosyllabic words, from 39 minimal pairs of the T2-T3 contrast. The identical targets were used because the study aimed to investigate the relationship between production and perception of the same target stimuli. Also, nasals, the only consonant category allowed in the coda position in Mandarin, were not used in the coda position due to the fact that they affect the vowel quality in many languages, including Mandarin. The target list consisted of 156 items: each of the 78 targets was repeated once.

A female native speaker of Mandarin was recorded in a quiet room reading the target list, in which the items were randomized. The reason that a female speaker was recorded for producing stimuli is that male speakers usually have lower pitch range than females, which might affect the perception of T3, which is known for its very low pitch

(creaky voice). All the stimuli were presented in Chinese characters as all native speakers of Mandarin have internalized the tones of most of the commonly used characters. This method of stimuli presentation is different from how the stimuli were presented to the L2 learners in the production task, in which the stimuli were typed with tone marks to give the L2 learners as much tonal information as possible.

After the recording, noise reduction was conducted using *Audacity* for the recording. Individual syllables were then extracted from the recording and organized on *Praat*. The two native speakers of Mandarin who served as the native judges in the production task also listened to the produced stimuli. One perceived 100% of the targets as the intended tones while the other native speaker identified over 98% of the targets as intended. Therefore, we can assume that the stimuli used in the current task are target-like and unbiased.

4.1.3 Procedure. The participants completed the identification test by using a script on *Praat*. Each L2 participant listened to the individual stimuli produced by the native speaker and was instructed to identify what tone they think they heard by choosing one of the tone options that appeared on the screen. Each participant completed two sessions in which the procedures and stimuli were identical to each other. The only difference was the tone options shown on the screen. The participants were presented with four tone options in the first session whereas in the second session, the participants chose from only two options, T2 and T3. During each session, each of the stimuli was presented twice, giving 156 items, in a randomized order for the L2 learners.

4.1.4 Analysis. As we recall the two analyses of the production task, Analysis 1 focuses on the stimuli that were not overtly perceived and does not categorize participants based on whether they are credited with having acquired the contrast. On the other hand, Analysis 2 enables the discussion of participants who were considered to have an overt contrast, a covert contrast, or lastly no contrast. Since the perception data were analyzed based on whether a L2 participant implements the T2-T3 contrast overtly or covertly, only the results from Analysis 2 in the production task were used as the criterion for grouping in the perception task.

There are three predictor variables in the current task; one is determined by the types of contrast that was implemented by the L2 participants, namely an overt contrast or a covert one. The second variable is the tone: T2, T3, or all targets (T2 and T3 combined). And the third variable is the session, which differed in the number of the options given to the participants, respectively four tones or two tones. Two outcome variables serve as the indicator of the participants' ability to identify the targets: the percentage of correctly perceived targets and the reaction time (RT). The reaction time was collected separately for the targets that were correctly identified and the ones that were not. Only the RT for the correctly identified stimuli were included in the analysis. Different from the production analyses, which primarily examined individual data for each L2 participant, the identification task here analyzed the aggregated data for testing the hypothesis in (5). Thus, the statistical analysis was constructed to compare the means among the participant groups. Three-way ANOVA was employed to complete the following comparison: 1) whether there is a main effect of participant groups in the

overall percentage of correctly perceiving the target tones and the reaction time, 2) whether there is a main effect of the target tones in the outcome variables, 3) whether there is a main effect of the tones provided to the participants to choose from, and 4) whether there is an interaction among the predictor variables on correctly perceiving the tones and on the reaction time.

4.2 Results

Table 4.1: The mean percentage of correct identification and the reaction time of the correct and incorrect identified targets, presented in the participant group, the target tones and the options provided to choose from

Participant Group	Target tones	Options to choose from	Percentage of correct identification of the targets	Reaction Time: correctly identified targets	Reaction Time: incorrectly identified targets
L2 Overt contrast (7 participants)	T2	4 tones	92%	1.855864345	2.142802427
		2 tones	97%	1.420366407	1.649599903
	T3	4 tones	33%	2.413306089	2.258402004
		2 tones	98%	1.471249373	1.385186264
	Both	4 tones	62%	1.967037229	2.285094647
		2 tones	98%	1.445077159	1.515119231
L2 Covert contrast (4 participants)	T2	4 tones	74%	2.018708068	2.171651735
		2 tones	75%	1.502391975	2.026475581
	T3	4 tones	18%	1.940841594	2.305430731
		2 tones	74%	1.591049006	2.009937445
	Both	4 tones	46%	2.049751824	2.272600985
		2 tones	74%	1.5543077	2.045662508
The participant who showed no contrast.	T2	4 tones	64%	2.966712918	3.706012658
		2 tones	93%	2.438879152	3.881468813
	T3	4 tones	2%	3.950137019	2.468647069
		2 tones	96%	2.612711258	4.64324371
	Both	4 tones	33%	3.004536922	2.801783958
		2 tones	95%	2.527569002	4.135393778

4.2.1 Percentage of Correct Identification. In this section, we first discuss the results on the percentage of correct identification of T2 and T3 from the speaker groups. Then we will examine the reaction time for the comparison. Table 4.1 below reports the mean percentage and the mean reaction time of each participant group in identifying T2,

T3, and T2 and T3 together. Data were reported separately for the two sessions, when the participants were given four or two options to choose from. The first column of the table lists the two participant groups as well as the only participant with no contrast at the bottom. The target tones are listed in the second column for each participant group. The third column is the options that the participants were given in the sessions, followed by the two outcome variables in the last three columns, namely the percentage of correct identification of the targets, and the reaction time for the correctly identified targets and the falsely identified ones.

The data of the percentage of correct identification is shown in Table 4.1 are presented graphically in Figure 4.1 and 4.2. We first examine the percentage of correct identification of the targets from the two participant groups when they were provided with four tone options to choose from, as shown visually in Figure 4.1. The participant groups are labeled by the bars of different patterns. From the first two bar clusters, it is shown that both groups identified T2 more correctly than T3, and the participant who has no contrast also shows a similar pattern. However, this finding does not hold true when the identifying options of tone were limited to two for the participants. Figure 4.2 shows the mean percentage of T2 and T3 targets correctly identified by the participant groups. We can see that when the options were limited to two for the participant to choose from, the percentage of the T2 that were correctly identified are approximately equal to that of the correctly identified T3. Such findings are supported by the statistic results obtained from the 3-way ANOVA test, which is presented in Table 4.2. It is indicated that there is a main effect of tone with a p-value close to 0. Also, there is a significant interaction found between the option condition and tone with a p-value close

to 0 as well. It indicates that the percentage of correctly identifiying T2 and T3 changes significantly when different options of provided, as shown in Figure 4.3.

Figure 4.1: The mean **percentage of correct identification** of targets from each participant group when given 4 tones to choose.

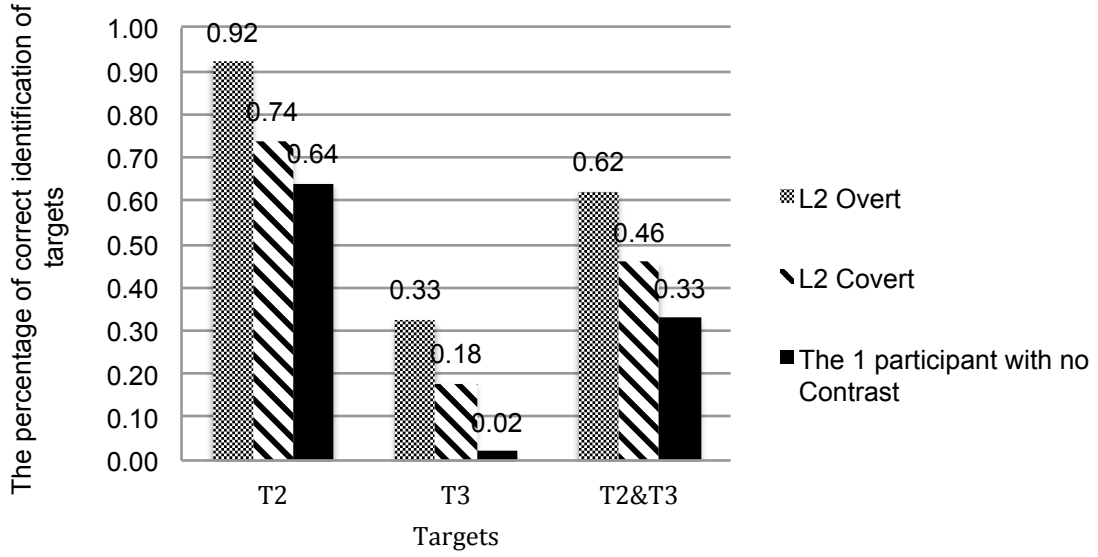


Figure 4.2: The mean **percentage of correct identification** of targets from each participant group when given only 2 tones to choose.

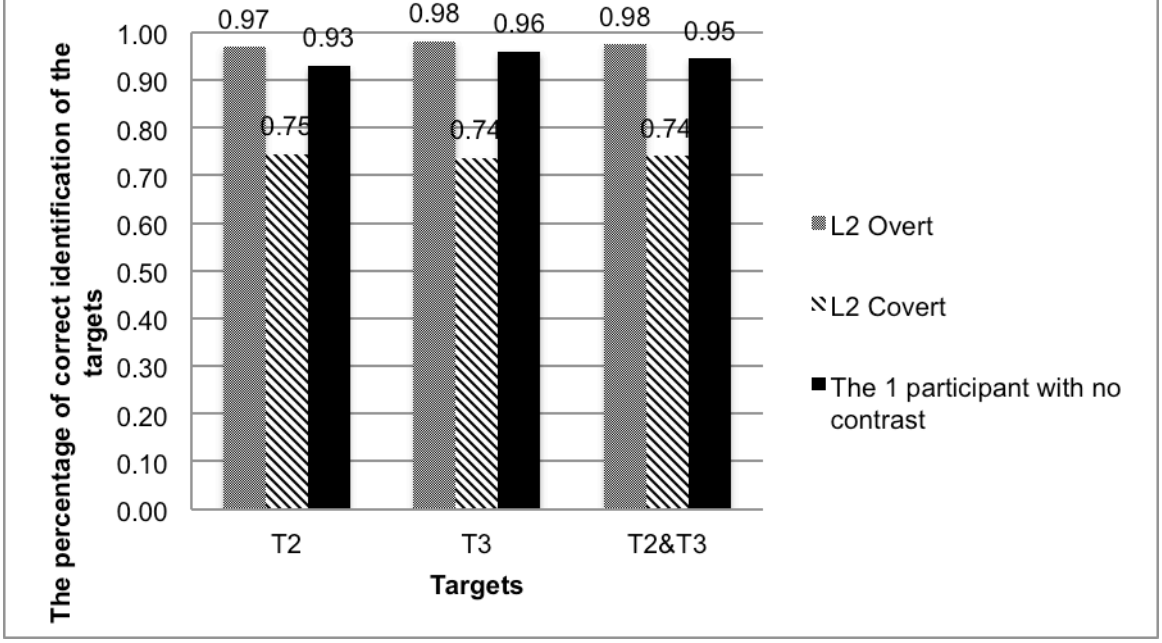


Table 4.2 Three-way ANOVA results of the percentage of correct identification (Speaker Group x Tones x Option Conditions)

Error: Speaker

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SpeakerGroup	1	0.4046	0.4046	2.323	0.162
Residuals	9	1.5671	0.1741		

Error: Speaker:Tone

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tone	1	0.9184	0.9184	51.168	5.35e-05 ***
SpeakerGroup:Tone	1	0.0001	0.0001	0.008	0.93
Residuals	9	0.1615	0.0179		

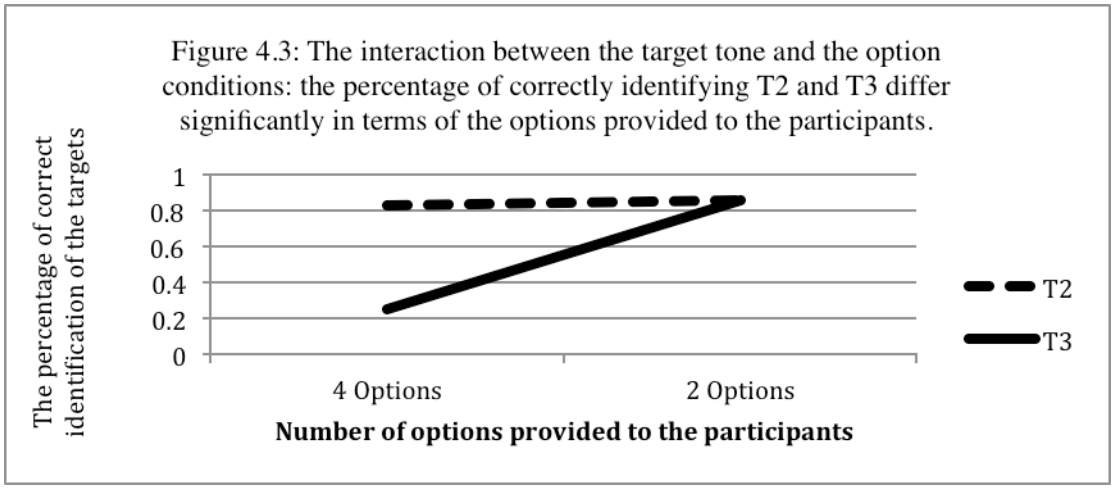
Error: Speaker:Options

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Options	1	1.1733	1.1733	22.983	0.000982 ***
SpeakerGroup:Options	1	0.0126	0.0126	0.246	0.631659
Residuals	9	0.4595	0.0511		

Error: Speaker:Tone:Options

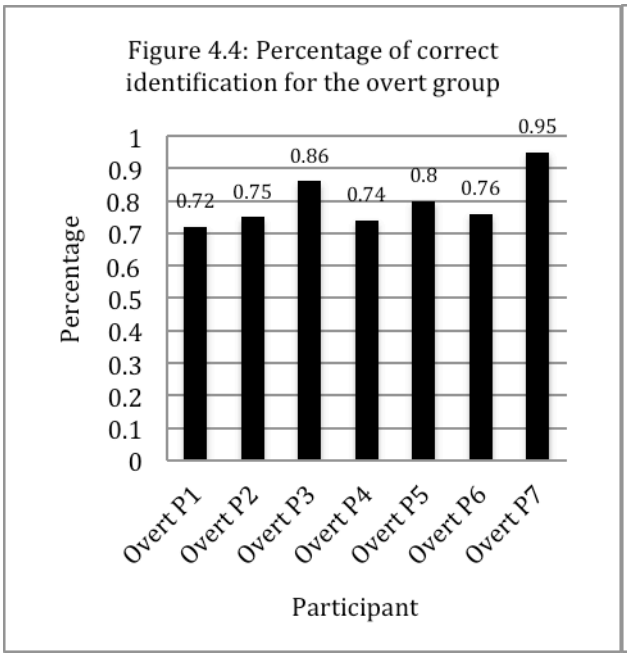
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Options:Tone	1	0.9476	0.9476	48.956	6.35e-05 ***
SpeakerGroup:Options:Tone	1	0.0018	0.0018	0.092	0.768
Residuals	9	0.1742	0.0194		

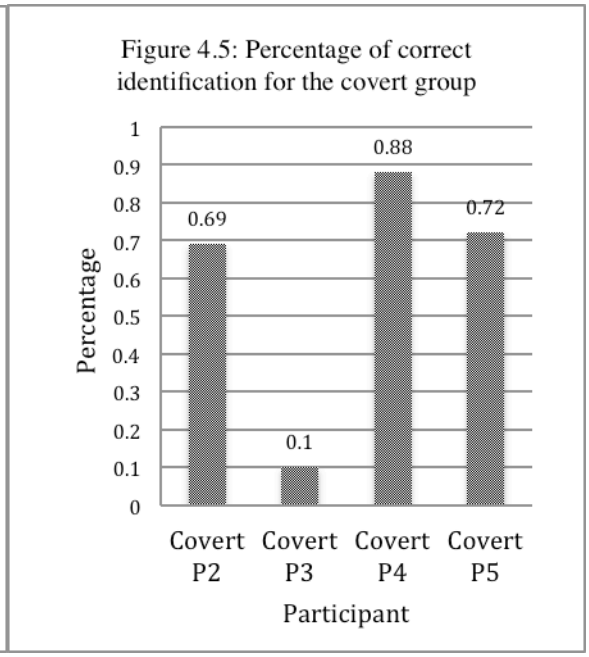
Signif. codes: 0 '***'



Secondly, we compare the two speaker groups in the percentage of correct identification in the four-option condition. From the first bar cluster in Figure 4.1, the overt group correctly identified 92% of T2 while the covert group identified 74% of the

T2. For T3, 32.6% were correctly identified by the overt group and 18% by the covert group. Similarly, from Figure 4.2, in the two-option condition, the overt group correctly identified 97% of the T2 and 98% of the T3 while the numbers by the covert group only falls at 74.5% for T2 and 73.7% for T3. For both option situations, and for both tones, a tendency has been shown; the overt group has correctly identified both the T2 and T3 better than the covert group. However, such a difference found between the two speaker groups is not statistically reliable with a p-value of 0.162.

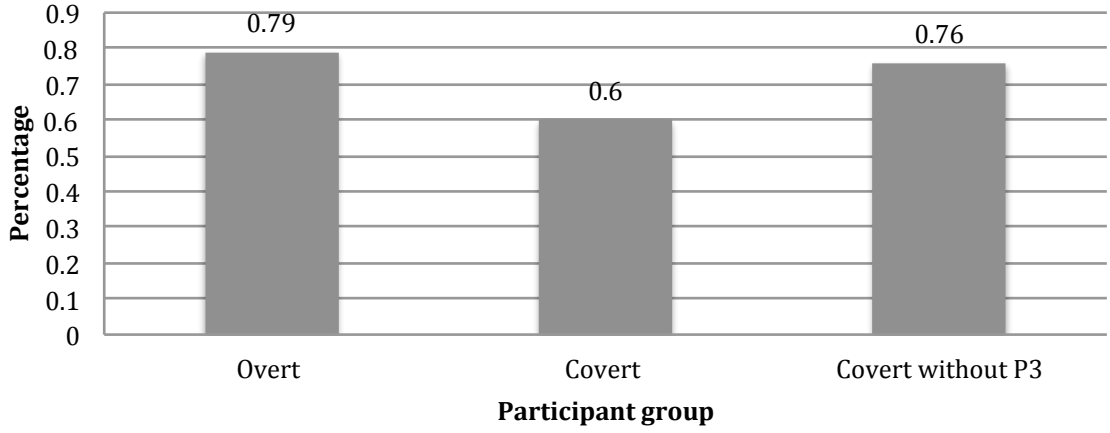




The statistically insignificant results can be explained by pointing out one of the participants from the covert group. Figure 4.4 and 4.5 show the percentage of correct identification for the two groups across the target tones and across the option conditions. Notice that the variation of the speakers in the overt group is small, but the speakers in the covert group varies in the percentage of correct identification. Specifically, Participant 3 (P3) from the covert group had an extremely low percentage of correct identification (10%) while the other participants in the same group have correctly

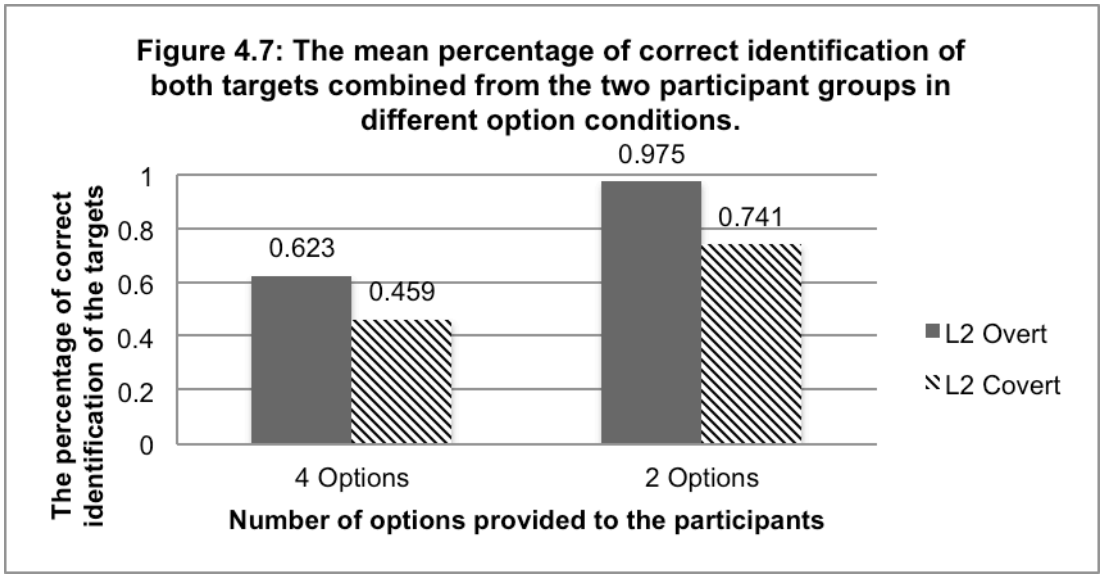
identified 60% to 90% of the targets. When including P3's data, the covert group has correctly identified 60% of the targets, which is 19% lower than the overt group's 79%. If we exclude P3 in the analysis, the covert group has correctly identified 76% of the targets which is only 3% lower than the overt group. The above comparisons were illustrated in Figure 4.6, in which the percentage of correct identification of the overt group, the covert group including P3 and excluding P3's data. The average P3's low percentage of correct identification has greatly lowered that of the covert group, which has a percentage that is seemingly a lot lower than the overt group. Neverthless, the trend that the overt group correctly identified the target tones better than the covert group still holds true while the difference between the two groups is not statistically significat.

Figure 4.6: Average percentage of correct identification of the overt group, covert group, and the covert group with P3's data



Now, we turn to investigate the main effect of the option conditions. Figure 4.7 below shows the percentage of correct identification of both target tones combined for each of the participant groups across the option conditions. From the figure, we see that

the percentage of correct identification of targets is 63% for the overt group in the four-option condition, and it has increased by 34.5% in the two-option condition to 97.5%. A similar increase has been found from the covert group of participants, who have correctly identified 46% of the targets in the four-option condition and 74% of the targets in the two-option condition. Such an increase from the four-option condition to the two-option condition was found to be statistically reliable with a p-value close to 0.



Lastly, we turn to the interaction between the participant groups and the other two predictor variables: tones and options. No significant interaction is found between the tones and the speaker groups with a p-value of 0.93. The difference between the percentage of correct identification of T2 and that of T3 does not change significantly across the speaker groups. Figure 4.8 below shows the relationship between the speaker groups and the tones. Also, the relationship between the speaker group and the options was not statistically reliable (*p=0.63*), which means that the percentage of correct identification does not change across the speaker groups when the number of the options provided changes. Figure 4.9 demonstrates such a relationship. Lastly, no

interaction was found among the three predictor variables on the percentage of correct identification.

Figure 4.8: The relationship between the speaker groups and the tones: The difference between the percentage of correct identification of T2 and that of T3 does not change significantly across the speaker groups.

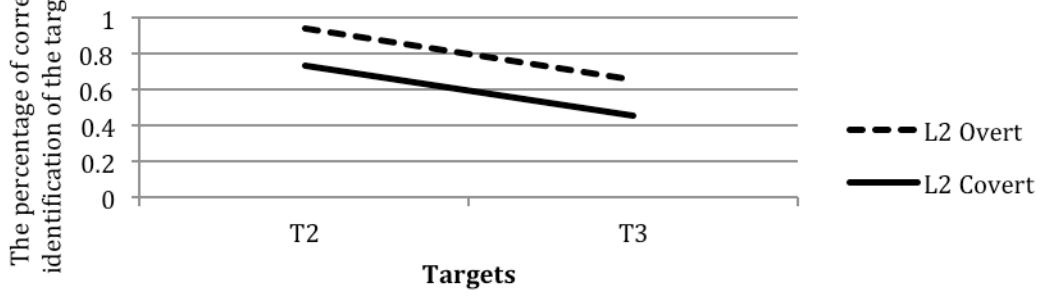
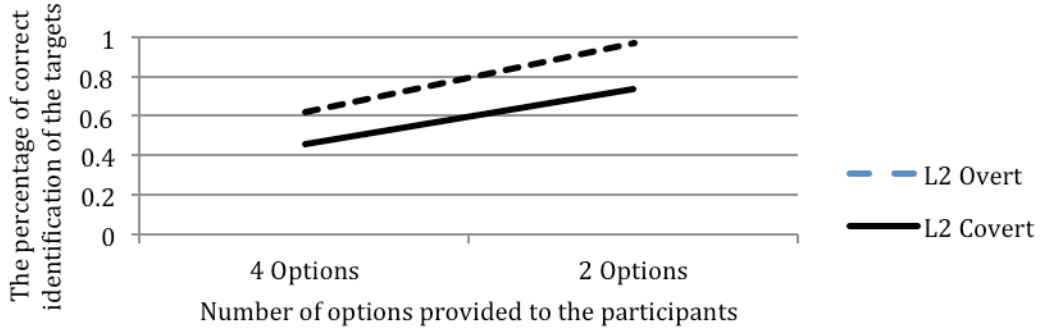


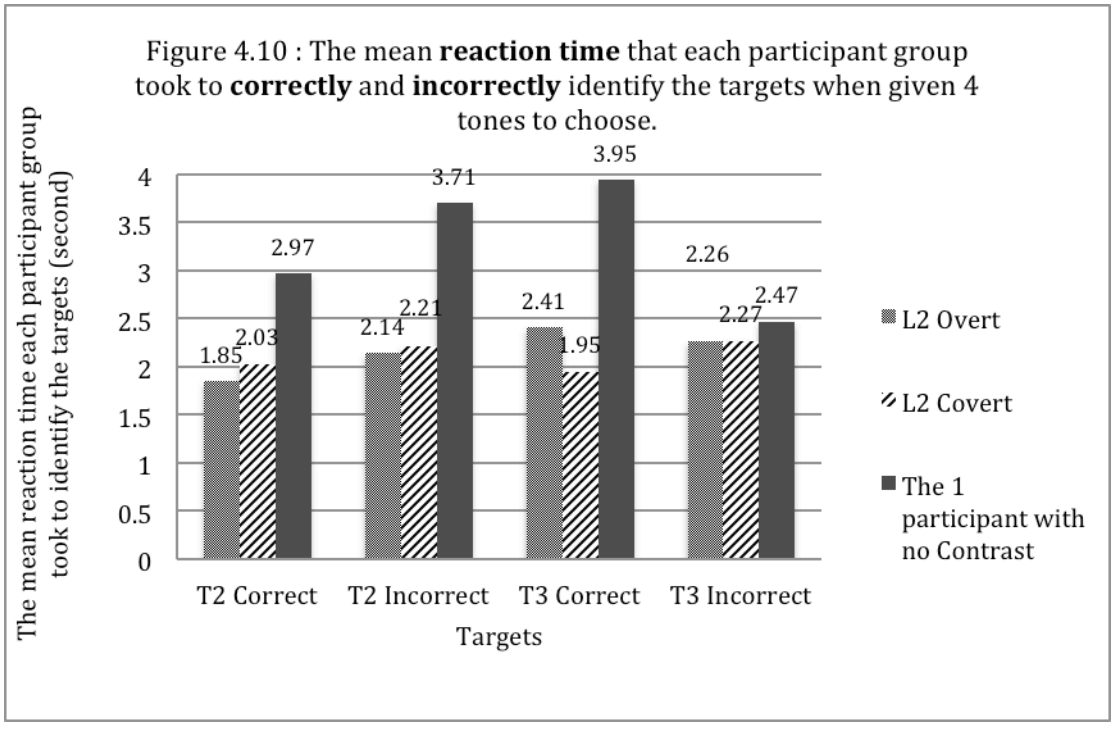
Figure 4.9: The relationship between the speaker groups and the option conditions: the percentage of correct identification does not change across the speaker groups when the number of the options provided changes.



To briefly summarize the results of the percentage of correct identification, there is a main effect of tone, as well as option condition. The effect of speaker groups is not significant. Regarding the interaction among the predictors, it has been shown that there is a statistically reliable interaction between the tone and the option condition, but no other interaction was found to be significant.

4.2.2. Reaction Time

Now, let us turn to the outcome variable, reaction time (RT), and how RT varies across the conditions of the three predictor variables. In this analysis, we have excluded the data of the participant who has been categorized in the covert group but showed a very different pattern from the other members of the group. The RT was computed separately for the items that were correctly identified and incorrectly identified. Figure 4.10 and 4.11 each illustrates the mean reaction time the correctly and incorrectly identified T2 and T3 by each group, respectively in the 4-option condition and the 2-option condition. The bars of different patterns represent the two speaker groups and the one participant who made no contrast, and the x-axis lists the correct T2, incorrect T2, correct T3 and incorrect T3.

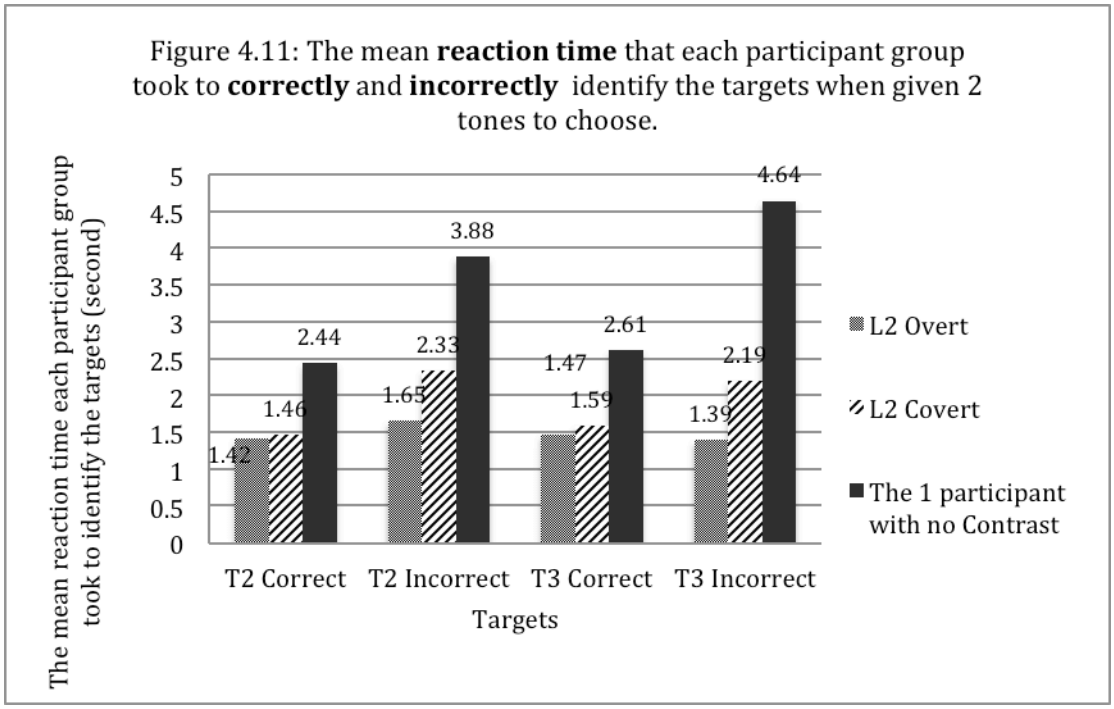


First, we examine Figure 4.10, which shows the results of the four-option condition. In general, the overt group shows a mean RT that is a little shorter than the covert group except for the correctly identified T3. When identifying the T3 target

correctly, the overt group took about 0.5 second longer in average than the covert group. Comparing identifying T2 and T3, the overt group took a little longer when listening to T3 targets. The covert group did not show a clear pattern in T2 and T3, but it took them less time to identify targets correctly than incorrectly. However, the differences observed and reported from Figure 4.10 seem fairly small. I will present the results of the significance test later to decide whether they are sufficient enough to be considered significant. As for the one participant who did not show any acoustic contrast, her mean RT is noticeably higher than the two L2 groups, regardless of the target tone and whether the targets were correctly identified.

Second, we now look at the RT that the participants took to identify the target tones when given 2 tones to choose from. Comparing the two groups, it is observed that the covert group shows a longer RT than the overt group in both identifying T2 and T3 correctly and incorrectly. On the one hand, both groups took longer when T2 tones were incorrectly identified than correctly perceived. On the other hand, the groups show different pattern in the identification of T3. The covert group still took less time in the correct identification than in incorrect response; however, the overt group reacted to T3 faster when the response was correct than incorrect. Regarding the differences between perceiving T2 and T3, both groups failed to show a clear pattern. When the tones were correctly identified, the overt group spent an average of 1.42 second on T2, which is 0.05 second faster than on T3 (1.47 second). The covert group also shows a faster RT for T2 than for T3 in correct identification. However, when the target is incorrectly identified, both groups have a longer RT for T2 than for T3 targets. Specifically, the overt group demonstrated a mean RT of 1.65 seconds for T2 targets,

which is 0.26 seconds slower than T3 targets. Similarly, the covert group has a longer mean RT to T2 than to T3.



Now we turn to the results of the significance test. Only the correctly identified items were being analyzed in the ANOVA test, for the following two reasons: 1) the interpretation of the incorrect responses is unclear itself, and 2) it would further complicate the already-complex statistical model with three predictor variables. The hypothesis the test was deigned to test concerns whether the two groups differ significantly from each other in identifying the targets. ANOVA shows that the effect of speaker group is not statistically reliable with a p-value of 0.872, which suggests that the reaction times of the two groups are not significantly different.

Table 4.3 Three-way ANOVA (Speaker Group x Tones x Option Conditions) results of the reaction time (RT)

Error: Speaker

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SpeakerGroup	1	0.011	0.0110	0.028	0.872
Residuals	8	3.182	0.3978		

Error: Speaker:Tone

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tone	1	0.4844	0.4844	4.949	0.0568
SpeakerGroup:Tone	1	0.1650	0.1650	1.685	0.2304
Residuals	8	0.7830	0.0979		

Error: Speaker:Options

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Options	1	3.838	3.838	18.831	0.00248 **
SpeakerGroup:Options	1	0.112	0.112	0.549	0.47990
Residuals	8	1.631	0.204		

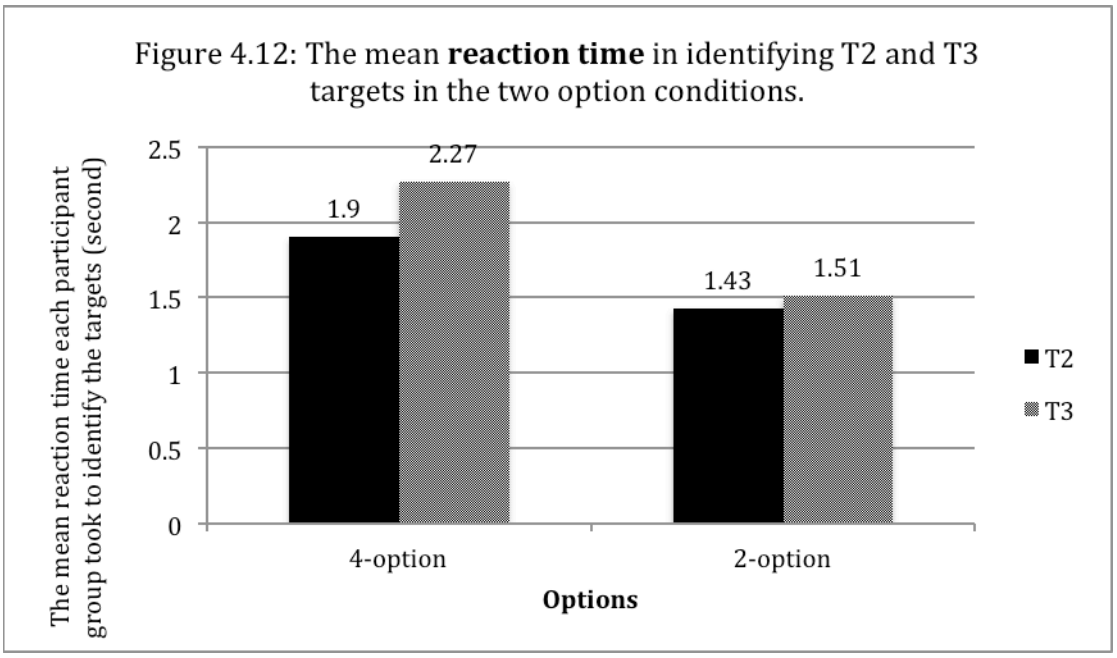
Error: Speaker:Tone:Options

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Options:Tone	1	0.2098	0.2098	1.328	0.282
SpeakerGroup:Options:Tone	1	0.2743	0.2743	1.736	0.224
Residuals	8	1.2641	0.1580		

Signif. codes: 0.001 '**'

The significance test also reveals an insignificant, yet near-significant, effect of the tone with a p-value of 0.056. Figure 4.12 visualizes the reaction time in identifying T2 and T3 regardless of the speaker groups in the two option conditions. As seen, the correct T2 targets required less time than the correctly identified T3 in both conditions. However, the RT difference between identifying T2 and T3 is not statistically reliable. Next, a main effect of the option condition has been shown, suggesting that the average RT in the two option conditions are reliably different from each other. Specifically, Figure 4.12 indicates that the participants spent a significantly longer time in the 4-option condition than in the 2-option condition. This result was expected because a smaller

number of options available limited the distraction of the participants and generated a faster reaction in correctly identifying the targets. Aside from the main effect of option conditions, no interaction among the predictor variables was found.



4.2.3 Interim Summary

This section summarizes the findings of the identification test. First of all, from the percent of correct identification, both groups of participants correctly identified T2 targets statistically better than the T3 targets. There is also a significant main effect of the option condition, which indicates that the participants' correct identification is statistically better in the 2-option condition than in the 4-option condition, particularly in identifying the T3 targets. There is a trend that the overt group identified the targets at a higher percentage than the covert group; nevertheless, the discrepancy between the two groups is not statistically reliable. An interaction was found between the tone targets and the option conditions, suggesting that the difference between correctly identified T2

and that of T3 changes significantly in the two option condition. Specifically, the participants correctly identified T3 reliably better in the 2-option conditions than in the 4-option condition, while the correct identification of T2 does not vary across the two conditions.

Secondly, from the results of the reaction time (RT), there is a main effect of the option conditions, which shows that participants correctly identified the targets within a significantly shorter time when only two options were available than when given four options. No other main effect was revealed from the statistical test on RT; however, the effect of tone was tested at a near significant confidence level. The results also do not indicate any reliable interaction among the three variables.

Chapter 5

Discrimination Task

5.1 Introduction and Hypotheses

As reviewed in Chapter 2, several studies have shown that L2 learners can perceive an L2 contrast before they can produce it. Under this assumption, it is probable that the L2 participants who produce the T2-T3 contrast covertly can perceive the contrast at a degree that is similar to the L2 participants who produce the contrast overtly. From the identification task, no significant difference has been found between the overt group and the covert group, which suggests that the overt group and the covert group have similar perceptual ability in identifying T2 and T3 targets. I conducted a perceptual discrimination task in order to further investigate the participants' perception in discriminating the two target tones. Pitch-manipulated productions of T2 and T3 were used as the perception stimuli, which differ systematically on the continuum of the acoustic measures.

The discrimination task aims to look for the subtle difference in perceiving T2 and T3 between the overt group and the covert group. That is, its goal is to investigate whether one group of the participants is more sensitive to the subtle difference in the pitch changes on the T2 and T3 continuum than the other participant group. There are two possibilities for the results. The first one is that the L2 participants who implemented the contrast overtly can distinguish stimuli with fine-grained pitch differences statistically better than the covert group. The rationale behind this possibility is the assumption that the participants who implemented an overt contrast may be more tuned in to the subtle

pitch difference, as it is the primary acoustic distinction between T2 and T3 productions. Another possibility is that the L2 participants who implemented the contrast covertly can distinguish stimuli with fine-grained pitch differences statistically better than the overt group. The reason that such a result is possible is that according to categorical perception, the target-like tone categories the overt group has established in their interlanguage which, in turn, enable them to produce target-like productions, might well hinder their ability to perceive the pitch difference that is not employed in the categorization of target-like tones in Mandarin.

Since both results are in principle plausible, the hypotheses that the discrimination task aims to test are formulated as below:

(6) The performance on the discrimination task by L2 participants who produce an overt contrast between T2 and T3 on the production task will not be reliably different from the performance on the discrimination task of the native speakers of Mandarin.

(7) The performance of L2 participants who implement a covert contrast will be systematically different from that of the native speakers of Mandarin.

5.2 Methodology

5.2.1 Participants. The same group of L2 participants participated in the discrimination task. In addition, there were also two control groups: a group of twenty-four native speakers of Mandarin and a group of five American English speakers who have no previous exposure to any tonal languages.

5.2.2 Stimuli. The non-pitch-modified stimuli recorded for the identification task were used as the baseline of the stimuli in the discrimination task. Specifically, the same 39 minimal pairs of T2 and T3 were systematically manipulated at two points in the pitch contour. The first manipulation is the F0 value at the turning point, which is the lowest in the entire contour, and the second is at the offset pitch height. On both acoustic measures, the two ends of the continuum are the F0 values of T2 and T3. The continuum was divided into three intervals, which makes four stimuli for each minimal pair (see Figure 5.1 below), where the stimulus with the highest F0 at both the turning point and offset pitch point is the values of an actual T2, and the stimulus with the lowest F0 at the measuring points was extracted from a T3. This procedure made the middle two stimuli the ones that are pitch-manipulated. In order to make the four stimuli equal in terms of the pitch manipulation, the stimuli at the two ends were also manually set at the pitch values measured from the actual T2 and T3. Since the F0 values at the contour onset for T2 and T3 are relatively close, which in turn do not serve as a contrasting cue for the two tones, the F0 values at the contour onset will not be included in the pitch manipulation. The stimuli were created by manipulating both of the T2 and T3 of each minimal pair, which served as the baselines in the pitch manipulation. That is, there were two sets of manipulated tokens for each minimal pair; one set was manipulated from the T2 as the baseline, whereas the other set was manipulated from the T3. Two sets of stimuli for each minimal pair were created to avoid any effect caused by the baseline stimuli.

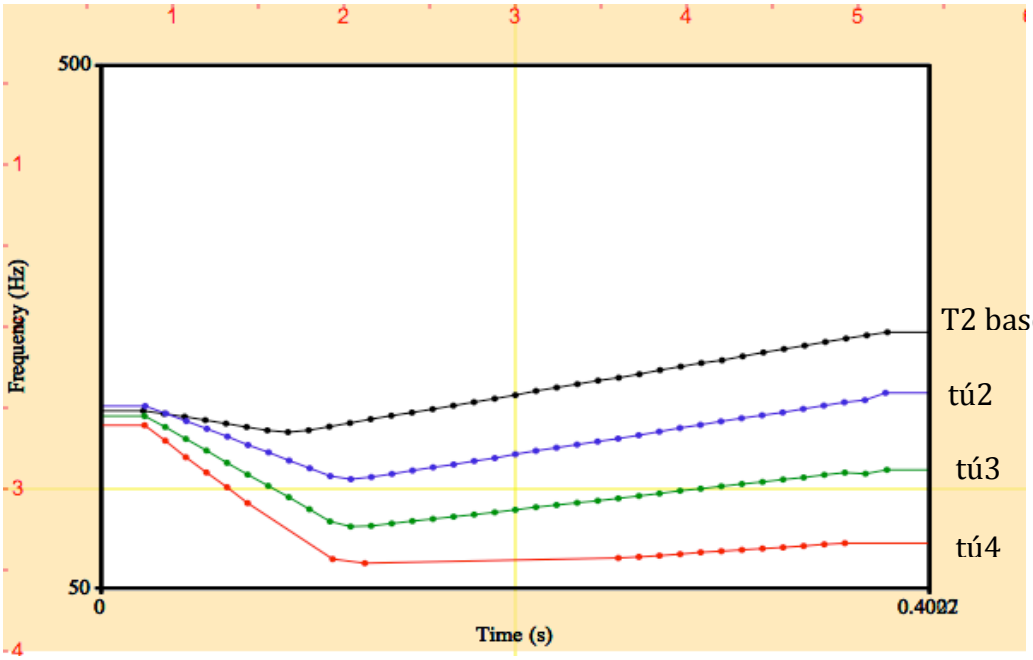


Figure 5.1 The four pitch-modified stimuli on the continuum of T2 and T3 in F0 at the turning point and offset.

For each minimal pair, the difference of the F0 value at the turning point between a T2 and a T3 (for instance, between *má* and *mǎ*) was divided into three intervals. The offset F0 was also adjusted by the same procedure. Then, when creating the set of stimuli from the T2 as the baseline, the pitch contour of *má* was manually manipulated, making another contour with a lowest F0 value that is a turning point interval lower than the lowest F0 of *má,* and an offset F0 that is an offset interval lower than the offset F0 of *má*. This procedure created the first manipulated pitch contour of each minimal pair. All stimuli were manipulated in the same way, by subtracting the corresponding interval to the F0 value at the turning point and the offset. The other set of the minimal pair used T3 as the baseline and was modified by adding a turning point interval to the lowest F0 of the T3, and by adding an offset interval to the offset F0 of the T3. All of the pitch

manipulation was performed on *Praat* by using the pitch manipulation function in which the pitch points between the onset and the turning point and between the turning point and the offset were removed. And then the F0 values at the turning point and offset were adjusted according to method described above. The baseline tones for each set were modified slightly in pitch by removing the pitch points between the onset and the turning point and between the turning point to the offset so that all stimuli of each same minimal pair systematically form a continuum. The manipulation procedure made four stimuli for each set of a minimal pair; *má1, má2, má3*, and *má4*, for the set with T2 as the baseline and *mă1, mă2, mă3, mă4* for the set in which the stimuli were modified based on the T3. The pitch-manipulated monosyllabic stimuli for each minimal pair were paired with one other stimulus created from the same set of the same minimal pair with an inter-stimulus interval of one second for the discrimination task. This makes 6 paired stimuli, respectively, *má1-má2, má2-má3, má3-má4, má1-má3, má2-má4*, and *má1-má4* for the set based on T2. Among the paired stimuli, the first three pairs are one interval apart whereas the fourth (*má1-má3)* and fifth (*má2-má4)* are two intervals apart, and the last one is separated (*má1-má4*) by three intervals. Figure 5.1 above shows the T2-T3 continuum of F0 value at the turning point and the offset. The figure was plotted on the basis of stimuli that were manipulated based on the T2 of */tu/,* where the black line is the T2 and red line is the T3 after manipulation. The distance between any two of the four lines is equal at the turning point and at the contour offset. With 39 minimal pairs, there would be 1248 paired stimuli in total. Specifically, there are 39 minimal pairs with 6 pairings of different stimuli for each minimal pair, with the orders

counter-balanced plus 4 pairings of the identical stimuli for each minimal pair, giving a subtotal of 624. There are two sets of pairings from the manipulation baselines, giving a grand total of 1248 tokens (39 x ((6 x 2) + 4) x 2 = 1248). In order to avoid fatigue during the task, only ten minimal pairs for a total of 320 stimuli were used for the discrimination task for all the participants (10 x ((6 x 2) + 4) x 2 = 320). The ten minimal pairs used in the discrimination task consisted of a stop consonant (/b/, /d/, /p/ and /t/) and a monophthong (/a/, /i/, /u/). The rational of choosing stops and excluding other consonants is that sonorants, such as nasals and laterals, are known for potentially affecting the quality of the following vowels. Stop consonants tend to have no effect on the following vowels. Also, if all 936 stimuli were to be included in the task, the perceptual sensitivity would likely decrease as the task gets excessively lengthy.

The discrimination task did not involve testing the sensitivity of the fine-grained acoustic difference on the durational cue that has been investigated in the production test. This is due to the large number of stimuli that are already included in the task. Also, the technical complexity involved in manipulating partial duration of a contour may impede gathering informative results. Testing the sensitivity on the durational cue will be a goal for future study.

5.2.3 Procedures. All participants listened to the paired stimuli and were instructed to indicate whether the two monosyllabic tones in each stimulus are the same or different. The task was completed on *Praat* by each of the participants, with the items being randomized for each participant by using a *Praat* script.

5.2.4 Analysis. There are three predictor variables in the discrimination task. The first one is the speakers' group; whether the participants are native speakers of Mandarin, American English speakers with no exposure to tonal languages, L2 learners who produce the target contrast overtly, or L2 learners who produce the contrast covertly. The second predictor variable is the number of interval between the two items in the paired stimuli. The distance between them is zero, one, two, or three intervals. That is, this predictor variable is how fine-grained an acoustic difference the participants can perceive. The last predictor variable, which is only for the purpose of avoiding baseline tone effect, is whether the stimuli are manipulated from the T2 baseline or T3 baseline of the minimal pair.

The outcome variable is the percentage of correct response. The average percentage of correct response was computed for each participant, when the stimuli are different intervals apart (4 interval distances: 0, 1, 2, or 3), and for each of the 2 baseline tones (T2 or T3). Therefore, we have eight (1 x 4 x 2 = 8) data points from each participant. Two analyses are presented; one for the items where the paired stimuli are acoustically identical (interval=0), and one for the items in which the stimuli are acoustically different (interval=1, 2, or 3). Analysis of Variance (ANOVA) was used to investigate the effect of each predictor variable and the interaction among the variables. Results of the items with identical stimuli are presented as Analysis 1, in which two-way ANOVA was perform with two predictors since data does not vary in terms of the interval. Analysis 2 reports the results of the items with different interval differences, and three-way ANOVA with an error term was employed in testing the significance of

the predictor effects. The error term in the ANOVA test accounts for the repeated measure that every participant has completed the discrimination of items of different intervals and different baseline tones. When ANOVA showed significance of a main effect or an interaction between the predictors, a post-hoc general linear model was built to reveal the occurrence of the significant effect. In both analyses, the major goal is to determine if there is a significant difference in the percentage of correct response among the participant groups, especially between the L2 overt group and L2 covert group. In addition, it is also the intention to see if the result of L2 overt contrast group resembles the participant group of native speaker and if the result of the L2 covert contrast group is more similar to the participant group of non-tonal language speakers. Similar to the identification task, we are interested in how the speaker groups differ from one another; therefore, aggregated data were analyzed.

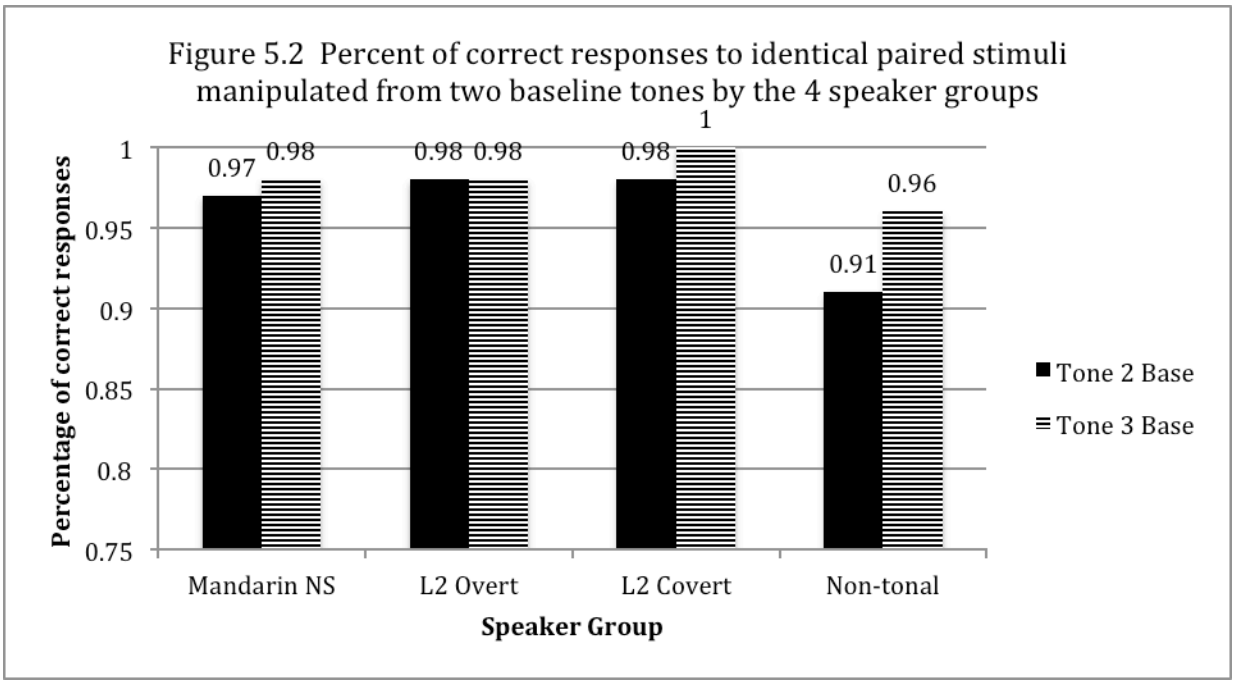
5.3 Analysis 1: Results of the items with identical stimuli

This section reports the results of different speaker groups in response to identical stimuli created from two baseline tones. Table 5.1 shows the average percent of correct response, which is also the percent of "same" response. And Figure 5.2 visualizes Table 5.1. From Table 5.1 and Figure 5.2, all four groups identified more than 90% of the items with paired identical stimuli correctly, as the same tones. Specifically, three groups, Mandarin Native, L2 Overt and L2 Covert, have shown very similar outcomes; each of them correctly perceived the pairs as the same at 97.5%, 98% and 99% of the time. The group of non-tonal language speakers showed a slightly lower

percentage of correct responses at 94%. Table 5.2 reports the outcome of the two-way ANOVA showing no significance in the effect of speaker groups with a p-value of 0.095.

Table 5.1 Percent of correct responses of different speaker groups in listening to identical stimuli from T2 and T3 baseline.

Participant group	Mandarin Native		L2 Overt contrast		L2 Covert contrast		Non-tonal Language Speakers	
Baseline Tone	**T2**	**T3**	**T2**	**T3**	**T2**	**T3**	**T2**	**T3**
% of "correct"	0.97	0.98	0.98	0.98	0.98	1	0.91	0.96
Overall % of "different"	0.975		0.98		0.99		0.94	



Now, let us examine the outcome separately for the two baseline tones. From Figure 5.2, we see that, while the L2 overt group demonstrated same percentages for both the T2-based and T3-based stimuli, the other three participant groups perceived the T2-based identical stimuli more correctly than the T3-based stimuli. On the one hand, the difference between the correct responses of T2-based and T3-based stimuli is

only 1% apart for the Mandarin NS, and 2% for the L2 covert group. On the other hand, the non-tonal speaker group showed a 5% difference between the two baseline sets with 91% of T2-based paired stimuli and 96% of T3-based ones perceived correctly. Regardless of speaker groups, the percent of correct response of T2-based stimuli is 96.7%, and that of the T3-based is 98.3%. As reported in Table 5.2, ANOVA shows a significant main effect of the baseline tones with a p-value of 0.0038, suggesting that the T3-based paired stimuli were perceived as the same tone significantly more than the T2-based ones.

Table 5.2 The Output of the two-way ANOVA (Speaker Group x Baseline tones)

Error: Speaker

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SpeakerGroup	3	0.01327	0.004424	2.287	0.0956
Residuals	35	0.06769	0.001934		

Error: Speaker:Baseline

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Baseline	1	0.005008	0.005008	9.598	0.00383 **
SpeakerGroup:Baseline	3	0.003943	0.001314	2.519	0.07392 .
Residuals	35	0.018262	0.000522		

Signif. codes: 0.001 '**'

Also, there is no significant interaction found between the speaker group and the baseline tone. This indicates that the difference observed in percent of correctness between the two baseline tones does not change significantly among the speaker groups. However, the significance value of the interaction is 0.07, which is approaching significance. This can be accounted for by the non-tonal speaker group's response. As discussed above, from outcomes obtained by the Mandarin NS, L2 overt and L2 covert, we expect a non-significant interaction between the baseline tones and the group.

However, the non-tonal speaker group shows that they perceived T2-based stimuli 5% less accurately than the T3-based stimuli. That is, the non-tonal group specifically demonstrates a different result from the other three groups in terms of perceiving the two baselines stimuli. Nevertheless, the speaker groups and the baseline tones do not significantly interact with each other.

5.4 Analysis 2: Results of discriminating stimuli that are different

This section reports the results of the four speaker groups in discriminating two acoustically different stimuli. The outcome variable is the same as in Analysis 1, the percentage of correct response; however its indication is different from Analysis 1 when we discussed the results of identical stimuli. The percentage of correct in Analysis 1 is the percentage of "same" response, whereas in Analysis 2, it is the "different" responses that are "correct" when the two perceived stimuli were acoustically different. In this case, what can be expected is that as the interval increases between the two stimuli, the easier the participants can discriminate them; thus, the percent of correct response increases as well. We will first discuss the results of T2-based stimuli, which are presented in Figure 5.3, and then proceed to the findings of T3-based stimuli, as illustrated in Figure 5.4.

In Figure 5.3 and 5.4, the four groups are listed on the x-axis and their average percentage of correct responses is presented in the three bars of different filled patterns. First, we observe the interval effect. From Figure 5.3, we see that all four groups show the results we expected, which is that as the greater the intervals are between the two

stimuli, the higher the percent of correct responses. Specifically, for the stimuli groups with one interval apart, Mandarin NS group shows the lowest percent of correct responses at 41% while the other three groups each show a percentage of correct responses that is between 56% to 60%. This suggests that among the four speaker groups, the Mandarin NS has the hardest time to discriminate two different stimuli when they are one interval apart, while the other three groups distinguished such stimuli pairs at chance level. When perceiving stimuli with two intervals apart, all groups showed above 80% of correctly discriminating the stimuli. When the stimuli are three intervals different, the non-tonal speaker group shows the highest percent (95%) of correctly discriminating the sounds with the L2 overt group showing the lowest at 86%.

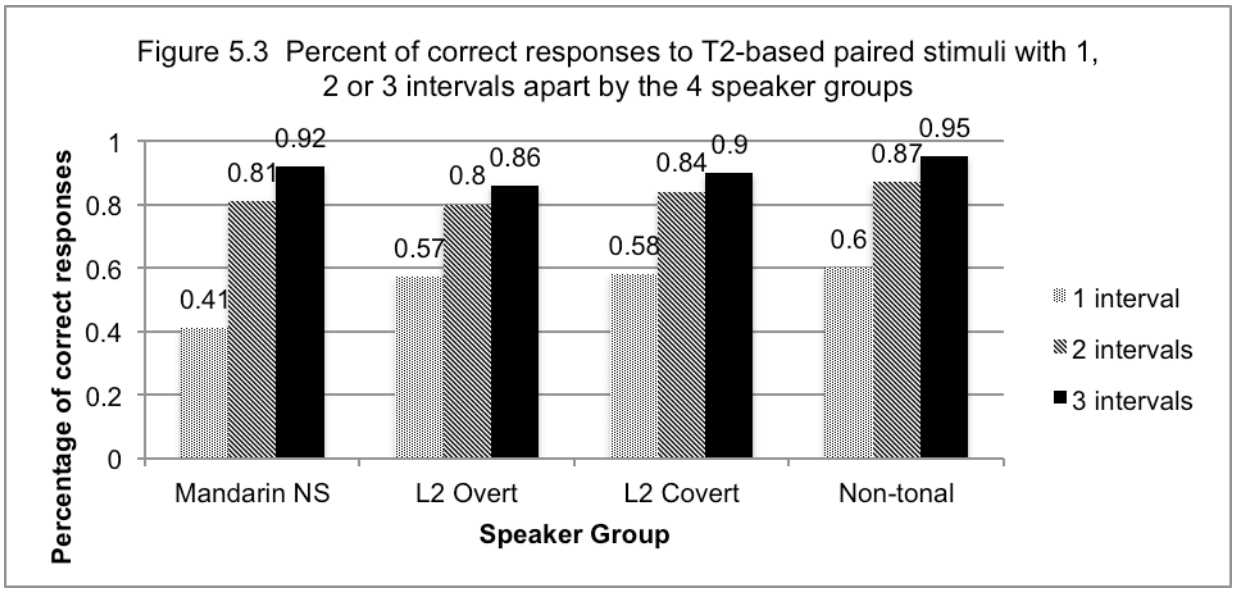


Figure 5.4 displays a very similar trend to that of Figure 5.3, in the way that the stimuli with one interval apart were least correctly discriminated and the ones with 3 intervals apart were most successfully distinguished for all four speaker groups. But the overall percent of correct responses to T3-based stimuli is lower than to T2-based

stimuli. This holds true for all four speaker groups and for all interval stimuli. For all three stimuli groups, the non-tonal speaker group shows the highest percentage of correct responses among the four stimuli groups, with the Mandarin NS group trailing behind. The two L2 groups show very similar patterns; they perceived 4% of the one-interval pairs, around 10% of the two-interval ones, and 17-20% of the three-interval stimuli as different.

Figure 5.4 Percent of correct responses to T3-based paired stimuli with 1, 2 or 3 intervals apart by the 4 speaker groups

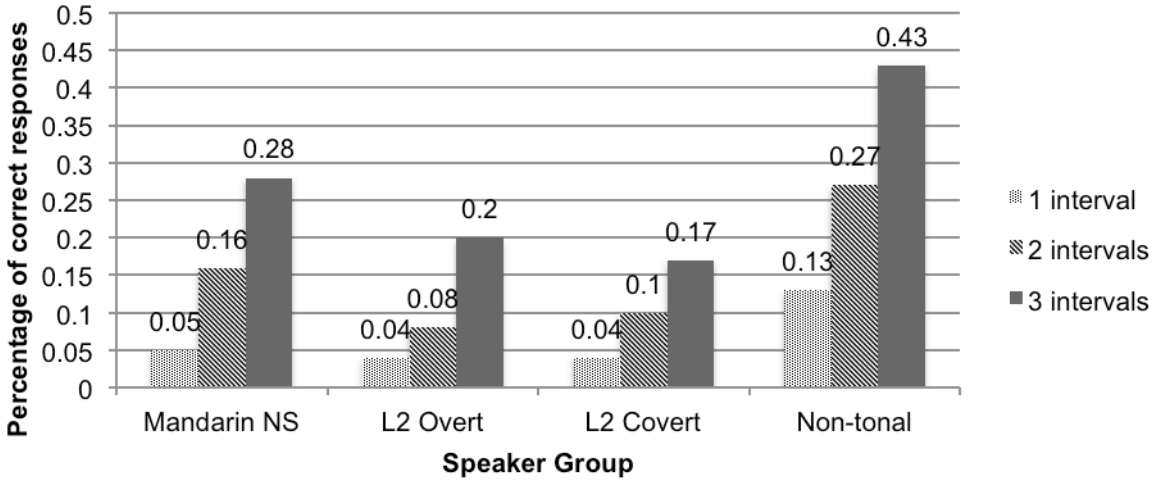


Table 5.3 reports the results of the three-way ANOVA test. First, it reveals a significant main effect of interval (*p=0*). A post-hoc general linear model, in which the results are reported in Table 5.4, indicates a statistically reliable difference between the one-interval and the two-interval stimuli, and between the one-interval and the three-interval, but the distinction between the two-interval stimuli and three-interval stimuli is not significant. This suggests that the amount of acoustical difference, F0 value at the turning point and the offset, in fact, reliably alters participants' ability to distinguish two tones, but the participants' perceptions do not differ reliably between when the two

tones are two intervals apart and three intervals apart. In order to illustrate what this means, Figure 5.5 was plotted for different intervals for the four speaker groups, regardless of the baseline tones. From the figure, we see that when the stimuli are one interval apart, the Mandarin NS participants considered them as different 23% of time with the other three groups considering them different at 30 to 37% of the attempts. When the stimuli are two intervals apart from each other, the percent of "different" response increase significantly to 49% for the Mandarin NS, 44% for the overt group, 47% for the covert group, and 57% for the non-tonal speakers. In discriminating tones that are three intervals different, the percentage of correctness increases non-significantly to 61% for the Mandarin NS group, around 53% for the L2 groups and around 70% for the non-tonal speaker group. This indicates that, although no categorical perception is clearly shown here as the percentage of correct does not reach the ceiling, from the significance test, it is evident that there is a "boarder" between the one-interval stimuli group and the two-interval stimuli group that statistically differentiates the participants' likelihood of actually perceiving two acoustically different tones as distinct. That is, when acoustic distinction is two or three intervals apart, the participants were more likely to perceive two tones as different.

Secondly, the significance test shows no significance in the effect of speaker groups, and the post-hoc test indicates no reliable difference is found between any of the two speaker groups in terms of correctly discriminating two acoustically different tones. Thirdly, the main effect of baseline tones is statistically reliable. Comparing Figure 5.3 and 5.4, we see an apparent difference between the T2-based and T3-based

sets. Participants distinguished T2-based stimuli statistically better than the T3-based ones. That is, T3-basd stimuli were perceived more frequently as the same tones by the participants.

Table 5.3 Output of the three-way ANOVA to test the significance of the predictor variables (Speaker Group x Baseline Tones x Interval)

Error: Speaker

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Speaker.Group	3	0.2431	0.08103	0.981	0.413
Residuals	35	2.8923	0.08264		

Error: Speaker:Basetone

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Basetone	1	19.353	19.353	655.688	<2e-16 ***
Speaker.Group:Basetone	3	0.129	0.043	1.462	0.242
Residuals	35	1.033	0.030		

Error: Speaker:Interval

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Interval	2	4.338	2.1689	257.371	< 2e-16 ***
Speaker.Group:Interval	6	0.178	0.0297	3.528	0.00416 **
Residuals	70	0.590	0.0084		

Error: Speaker:Basetone:Interval

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Basetone:Interval	2	0.7072	0.3536	34.357	4.02e-11 ***
Speaker.Group:Basetone:Interval	6	0.0625	0.0104	1.012	0.425
Residuals	70	0.7205	0.0103		

Signif. codes: 0 '***' 0.001 '**'

Table 5.4 Output of the post-hoc test on the main effect of intervals and the baseline tones.

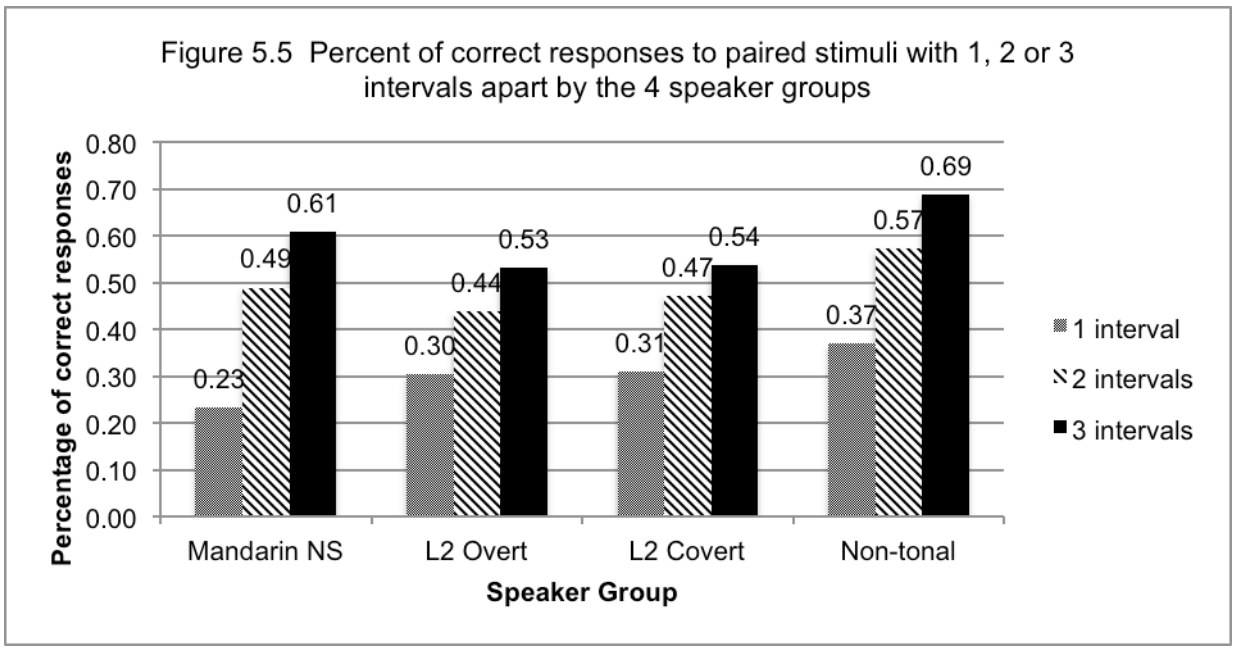
Linear Hypotheses:

		Estimate	Std. Error	z value	Pr(>|z|)
Interval 2 - Interval 1	0	0.27077	0.08183	3.309	0.00266 **
Interval 3 - Interval 1	0	0.34577	0.08183	4.226	< 1e-04 ***
Interval 3 - Interval 2	0	0.07500	0.08183	0.917	0.62979

Linear Hypotheses:

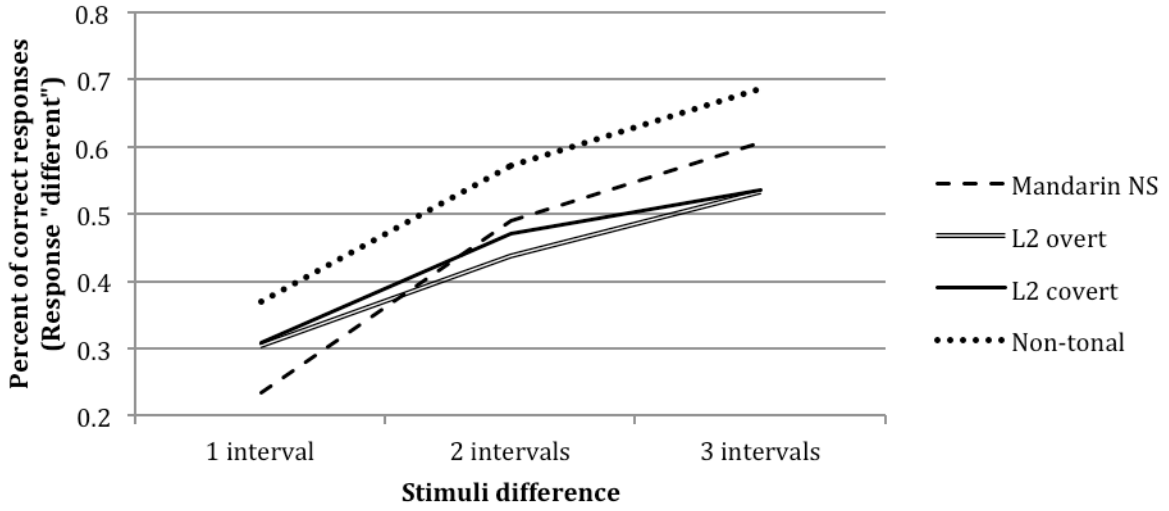
		Estimate	Std. Error	z value	Pr(>|z|)
T3 - T2	0	-0.47088	0.08183	-5.755	8.69e-09 ***

Signif. codes: 0 '***' 0.001 '**'



Next, let us examine the interactions between and among the three predictor variables. There is a significant interaction found between the speaker groups and the intervals (*p=0.00416*) from Table 5.3, suggesting that the difference found between two speaker groups changed significantly when discriminating stimuli of different intervals. Figure 5.6 below plots the interaction between the speaker groups and the interval. And Table 5.5 presents the results of the post-hoc test of the interaction between the speaker groups and the interval. The first half of Table 5.5 reports the significance of the difference between two intervals within each speaker group, the second half holds the same interval and tests the distinction between two speaker groups.

Figure 5.6 The interaction between the speaker groups and the intervals on percent of correct (different) responses



On the one hand, the post-hoc test shows no significant effect of speaker groups within each of the four intervals. On the other hand, there is some statistically reliable effect of interval within a certain speaker group. Table 5.5 reveals that the percent of correct response differs significantly between one-interval stimuli and three-interval stimuli within the Mandarin NS group, and such a significant distinction was also revealed for the L2 Overt group. In other words, the Mandarin NS group and the L2 Overt group resemble each other regarding the effect of interval, especially between one-interval and three-interval. It is interesting that the other two groups did not have such an effect as the interval main effect found between one-interval and three-interval is extremely reliable. One possible reason is that the participant number in the L2 covert group and the non-tonal speaker group is relatively small, which can lessen the significant value of the predictor effect.

Table 5.5 Output of the post-hoc test of the interaction between the speaker group and the interval

Within each speaker group, the difference among the intervals:

		Estimate	Std. Error	z value	Pr(>|z|)
Non-tonal	Interval 2 - Interval 1	-0.1355	0.1294	-1.047	0.8605
Non-tonal	Interval 3 - Interval 1	0.2031	0.1694	1.199	0.7757
Non-tonal	Interval 3 - Interval 2	0.3386	0.1294	2.616	0.0612
L2Covert	Interval 2 - Interval 1	-0.3787	0.1694	-2.236	0.1569
L2Covert	Interval 3 - Interval 1	-0.5542	0.2132	-2.600	0.0641
L2Covert	Interval 3 - Interval 2	-0.1754	0.2132	-0.823	0.9475
L2Overt	Interval 2 - Interval 1	0.3178	0.3453	0.920	0.9158
L2Overt	Interval 3 - Interval 1	0.5786	0.2067	2.800	0.0367 *
L2Overt	Interval 3 - Interval 2	0.2609	0.2494	1.046	0.8610
Mandarin NS	Interval 2 - Interval 1	0.3178	0.3453	0.920	0.9158
Mandarin NS	Interval 3 - Interval 1	0.5786	0.2067	2.800	0.0368 *
Mandarin NS	Interval 3 - Interval 2	0.2609	0.2494	1.046	0.8610

Within each interval, the difference among the speaker groups:

		Estimate	Std. Error	z value	Pr(>|z|)
Interval 1	L2Covert - EN	-0.060038	0.169420	-0.354	1.000
Interval 1	L2Overt - EN	-0.065216	0.150174	-0.434	1.000
Interval 1	NS - EN	-0.135475	0.129396	-1.047	0.942
Interval 1	L2Overt - L2Covert	-0.005179	0.150174	-0.034	1.000
Interval 1	NS - L2Covert	-0.075437	0.129396	-0.583	0.998
Interval 1	NS - L2Overt	-0.070259	0.102921	-0.683	0.995
Interval 2	L2Covert - EN	-0.034375	0.150174	-0.229	1.000
Interval 2	L2Overt - EN	0.438562	0.272304	1.611	0.644
Interval 2	NS - EN	0.431292	0.234551	1.839	0.482
Interval 2	L2Overt - L2Covert	0.472937	0.260766	1.814	0.499
Interval 2	NS - L2Covert	0.465667	0.221051	2.107	0.310
Interval 2	NS - L2Overt	-0.007271	0.182994	-0.040	1.000
Interval 3	L2Covert - EN	0.270804	0.329638	0.822	0.985
Interval 3	L2Overt - EN	0.120625	0.376118	0.321	1.000
Interval 3	NS - EN	0.260854	0.249379	1.046	0.943
Interval 3	L2Overt - L2Covert	-0.150179	0.212379	-0.707	0.994
Interval 3	NS - L2Covert	-0.009949	0.165337	-0.060	1.000
Interval 3	NS - L2Overt	0.140229	0.245234	0.572	0.999

Signif. codes: 0.01 '*'

Lastly, we turn to the interaction between the intervals and the baseline tones. From Table 5.3 in one of the earlier pages, the interaction between the intervals and baseline tones were revealed to be very robust (*p=0*). It suggests that the baseline tones have a statistically different effect depending on the intervals between the stimuli

in question for discrimination. Figure 5.7 illustrates the interaction between the two predictor variables. First of all, it shows the baseline tone main effect, that participants discriminated stimuli created from T2 better than T3. Secondly, the more intervals between the stimuli, the better the discrimination by the participants. When holding the interval set constant, it is seen that for all three stimuli, the percent of correct discrimination of the T2-based stimuli is much higher than that of T3-based. This is confirmed by the post-hoc test (reported in the first half of Table 5.6), which shows a statistically significant effect of the baseline tone within each interval set (*p=0*).

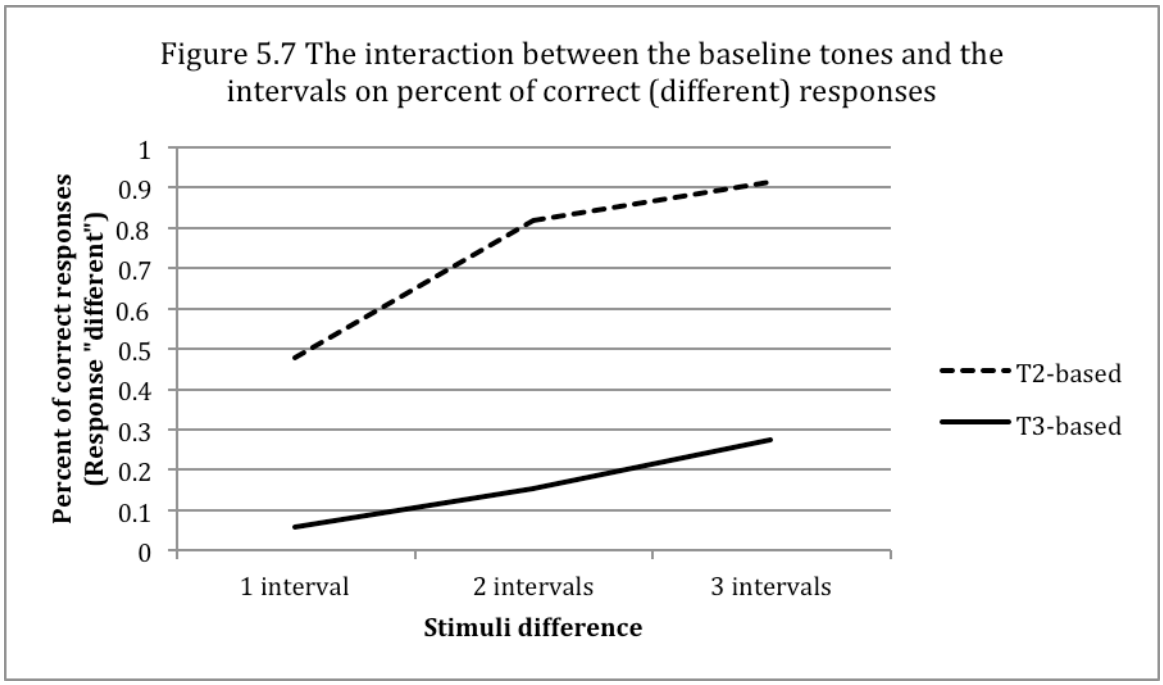


Table 5.6 Output of the post-hoc test of the interaction between the baseline tone and the interval

Within each interval, the difference among the baseline tones:

Linear Hypotheses:

	Estimate	Std. Error	z value	Pr(>|z|)
Interval 1:T3 - T2	0.43628	0.02706	16.12	<2e-16 ***
Interval 2:T3 - T2	0.76258	0.02706	28.18	<2e-16 ***
Interval 3:T3 - T2	-0.82873	0.05413	-15.31	<2e-16 ***

Within each baseline tone, the difference among the intervals:

	Estimate	Std. Error	z value	Pr(>|z|)
T2:Interval 2 - Interval 1	0.43628	0.02706	16.121	< 1e-06***
T2:Interval 3 - Interval 1	-0.42053	0.02706	-15.538	< 1e-06***
T2:Interval 3 - Interval 2	-0.85681	0.02706	-31.659	< 1e-06***
T3:Interval 2 - Interval 1	0.19027	0.03827	4.971	2.1e-06***
T3:Interval 3 - Interval 1	-0.63846	0.02706	-23.591	< 1e-06***
T3:Interval 3 - Interval 2	-0.82873	0.05413	-15.311	< 1e-06***

Signif. codes: 0 '***'
(Adjusted p values reported -- single-step method)

What the post-hoc test also shows is that within the T2-based stimuli group, the discrepancy between any two intervals on the percent of correct response is statistically significant (*p=0*). The same results were also found for the T3-based; the difference between any two intervals was confirmed reliable. No interaction among the three predictor variables was found.

5.5 Interim Summary

To conclude this chapter, I have briefly summarized the findings obtained from the discrimination task. First, there is a strong main effect of the baseline tones for both the acoustically identical stimuli and the different stimuli, which indicates that participants perceived T3-based stimuli more as the same than T2-based stimuli. No significant effect of the speaker groups was shown, suggesting that the four speaker

groups did not show a statistically reliable distinction among one another in correctly discriminating pitch-modified stimuli. However, for the stimuli that are acoustically different, the interaction between speaker groups and the interval was revealed significant. The significance occurs in the Mandarin NS group and the L2 Overt group, which have demonstrated a reliable distinction between stimuli that are one interval apart and ones that are three intervals apart. Such a distinction was not present in the results of the L2 covert group and the non-tonal language speaker group. In addition, for acoustically different stimuli, there is also a significant interaction between the baseline tones and the interval sets. It suggests that the difference of the participants' discrimination on the different interval sets is dependent on the baseline tones.

The next chapter presents a general discussion of the findings of this dissertation, and focuses on the discussion of the hypotheses proposed for the three experiments: the production tests, the identification task and the discrimination tasks.

Chapter 6

Discussion

This chapter presents a discussion of the results obtained from each of the three tasks and how these results help support the proposed hypothesis. I will attempt to account for these results from the perspectives of the L2 participants' interlanguage systems and the implications of previous studies on L2 acquisition. Taking both of the production analyses together, Section 6.1 primarily concerns the following three aspects from the production task: 1) the occurrence of a covert contrast, 2) the acoustic measures that were implemented in the T2 and T3 productions, and 3) the longitudinal progress that the L2 participants have made. Section 6.2 presents a general discussion on the identification task, in which the primary objective is to see if the L2 participants at different stages of T2-T3 production identified T2 and T3 differently. Section 6.3 lays out the implications of the discrimination tasks on the basis of the perceptual sensitivity to subtle acoustic details. Each of the sections will address each of the hypotheses proposed for the relevant task. Lastly, Section 6.4 discusses the general implications of the findings of the current study.

6.1 Production Results

6.1.1 The Occurrence of a Covert Contrast

One of the primary goals of this dissertation is to attest the occurrence of a covert contrast in the acquisition process of Mandarin T2 and T3 by L2 learners. In Chapter 3, I have presented two different analyses of the L2 data of T2 and T3 production, differing in the scoring criterion and the use of the 80% target-like threshold. Both of the

analyses have shown that in cases where the target tones were not perceived as target-like by the native transcribers, some of the L2 participants implemented a statistically reliable distinction between their T2 and T3 production on at least one of the acoustic measures used in the study. Specifically, Analysis 1 shows that four participants in the first elicitation, and three participants in the second elicitation implemented a minimum of one acoustic distinction that is statistically significant between the T2 and T3 targets that were not perceived as target-like by the native Mandarin transcribers. Analysis 2 reveals that in the first elicitation, all of the participants whose target-like productions did not reach 80% of all of the targets elicited made a statistically reliable distinction between their T2 and T3 productions, and in the second elicitation, all but one participant made such a distinction. Taking both analyses together, the findings suggests that the L2 participants have produced non-target-like T2 and T3 that in fact are reliably different from each other; in other words, they produced a covert contrast between T2 and T3.

The occurrence of a covert contrast can be accounted for from the perspective of the interlanguage system of the L2 participants. According to Selinker (1972), an interlanguage is a linguistic system that L2 learners construct in their mind, which is dependent upon the individual language background of each L2 learner, including their L1 and L2 experience. Since interlanguage is individually built in every learner's mind on the basis of an individual learner's language experience, it must be treated individually and independently from the target language. The occurrence of a covert contrast on the production task is evidenced as a stage in which the L2 participants' production of a contrast is not perceived as target-like but is systematically established

in their IL. The occurrence of such a contrast indicates that in a L2 learner's interlanguage, the difference between the phonemes of the target contrast has been established and reliably implemented in the relevant acoustic cues, but before the distinction reaches the point in which it can be perceived by the native speakers of the target language, the contrast is considered to remain not target-like. This finding points out one of the problematic aspects of the commonly used criterion to determine whether a L2 learner has acquired the production of an L2 phonemic contrast, which mainly depends on whether the contrast has been perceived by native speakers of the target language. Such a criterion fails to take the nature of L2 acquisition into account, that L2 acquisition is an ongoing process and that L2 learners might have implemented a distinction that is not yet audible to the native speakers. In many cases, a L2 learner who produces a covert contrast is not given credit for making a contrast because the perceptual ability of the native speakers of the target language has been sensitive only to the phonemic categories that are relevant to the target language, failing to recognize the covert contrast that might fall within a single category in the target language. Similar to how an interlanguage is viewed independently from the target language, how a L2 learner implements a contrast ought to be treated independently from whether the contrast is audible to the native speaker of the target language. As L2 acquisition is often assumed to be an ongoing process, a covert contrast produced by L2 learners in their interlanguage should be viewed as an essential stage that is between the stage of no contrast and a stage where a target-like contrast is implemented.

6.1.2 The Acoustic Measures Implemented in the Contrast

A statistically reliable contrast was implemented by the participants with different patterns of the three acoustic measures. That is, some participants showed a reliable contrast on only one of the three acoustic measures whereas some have implemented it on different combinations of two or all of the acoustic measures. From the first elicitation, Analysis 1 shows four participants making a covert contrast; two of them have shown it on two acoustic measures and the other two made such a contrast on one acoustic measure. What the four participants' non-target-like productions shared is that they all have demonstrated a reliable distinction on the duration between the contour onset and the turning point. A similar pattern was seen from Analysis 2 as well, where all participants, except for one, have made a statistically reliable distinction between T2 and T3 production on the duration cue, regardless of what other measure was also employed to implement the contrast. It suggests that from the first elicitation, in the cases that the T2 and T3 productions were not considered target-like, most participants have implemented the contrast covertly in the duration measure. This indicates that when acquiring tones, L2 learners might pay more attention to the time cue because it is easier for them to implement the contrast. This finding can be attributed to the L1 influence on the L2 learners' interlanguage phonology. Tarone (1976, 1978) and Sato (1983) have indicated that L1 phonological structures have showed effect on L2 syllable production. So (2006) also demonstrated that L1 prosodic background play a role the L2 production of Mandarin tones. Here, the finding that the L2 participants from the first elicitation have largely implemented a covert contrast between T2 and T3 on the durational cue can be explained by the temporal contrasts of consonants and vowels in their L1, American English. Some studies suggest that the temporal cue is employed in

the production and perception of American English vowel (Hillenbrand et al 1995), and VOT, also a temporal measure, is a primary acoustic feature of stop consonants in American English (Liberman et al, 1957; Lisker and Abramson, 1970). The participants tend to use the temporal cue in producing a non-native contrast in the first elicitation more often because they are more familiar with the temporal cue, which is widely used in making a phonemic contrast in the participants' L1. Given that it is widely used in the L1, we can assume that the temporal measure is easier to acquire than the pitch contour (F0 difference) and the fixed pitch point (offset pitch height). As a result, this suggests that the L1 might have an effect in producing Mandarin T2 and T3. However, Hao (2013) shows that L1 background does not necessarily play a role in tone production, especially for the contrast between T2 and T3. There might be other factors which influence the acquisition of T2-T3 contrast. The finding of the second elicitation might provide some support for some other aspects that can influence the acquisition of T2 and T3 production here by showing a different pattern of the acoustic measure employed in the implementation of a covert contrast. This is discussed below.

The second elicitation shows a different pattern in terms of the acoustic measures implemented by the L2 participants. From Analysis 1, three of the twelve participants in the second elicitation made a significant distinction between the non-target-like T2 and T3, but none of them implemented it on the duration cue. Instead, the acoustic measure that all three of them used to implement such a contrast is the F0 difference. Two of them have also implemented a reliable distinction on the offset pitch height. Analysis 2 shows that both the F0 difference and the duration cue were both employed by three participants (out of the five whose target-like production did not

reach 80%). The pattern observed in the first elicitation, that the majority of the participants implemented a covert contrast on the duration measure, was not found in the second elicitation, which was eighteen months after the first one. However, no indication can be drawn based on the different patterns observed here because of the following two reasons. First, in Analysis 1, the four participants who implemented a covert contrast in the first elicitation are not completely the same as the ones in the second elicitation. Among the four in the first elicitation, only one was found to implement such a contrast in the second elicitation. Her production will be discussed in detail later. Second, in Analysis 2, all of the participants that were not given credit for acquiring the T2-T3 contrast in the first elicitation have produced at least 80% target-like T2 and T3; thus, they were then credited with having acquired the contrast in the second elicitation. The participants without being credited with having acquired the contrast in the second elicitation were all later enrolled. Therefore, no comparison can be made to the different groups of participants on the acoustic measures used. Nevertheless, we can discuss the participant who has made a covert contrast in both of the elicitations from Analysis 1.

Table 6.1 puts together Participant FZT's results from the first and the second elicitation in Analysis 1, in terms of the acoustic measures that she made a covert contrast with the target-like percentages of T2 and T3. On the one hand, we see that in the first elicitation, she made a statistically reliable distinction between the non-target-like T2 and T3 on the duration to the turning point and the F0 difference. At that time, 57% of the T2 and 18% of the T3 were perceived as target-like. On the other hand, her production from the second elicitation reveals that she implemented a covert contrast on

the F0 difference and the offset pitch, but not on the duration measure, where 64% of T2 and 49% of T3 were identified as target-like by the native transcribers. Both elicitations show a covert contrast in the F0 difference, but as the target-like percentage significantly increased (shown in Chapter 3, Table 3.4) from the first to the second elicitation, especially target-like T3, the statistically reliable distinction found in her non-target-like production has shifted from the duration measure to another pitch measure. One possible way to account for this is that as this participant continued to receive exposure in Mandarin during the eighteen months between the two elicitation, she has acquired the offset pitch as one of the acoustic cues to produce contrasting T2 and T3. Her IL grammar has developed from being influenced by her L1 to an IL grammar that is more similar to the target language, in which pitch and F0 measures are essentially the primary characteristics of tones (Moore and Jongman, 1997).

Table 6.1: The acoustic measures in which Participant FZT's non-target-like productions show a covert contrast from Analysis 1.

Participant FZT	NT2 vs NT3			Percent of target-like T2	Percent of target-like T3
	F0 Diff	Duration	Offset		
First Elicitation	+	+	-	57%	18%
Second Elicitation	+	-	+	64%	49%

The above explanation might also be applicable to the similar pattern of the non-target-like productions observed from the other participants in the first and the second elicitation. The finding that all four participants with a covert contrast in the first elicitation have implemented such a contrast in the duration measure while all three in the second elicitation showed it in the F0 difference may suggest that when participants made a covert contrast between the non-target-like tones, they may use the duration as a primary measure before acquiring the F0 contrast when receiving sufficient exposure

in the target language. It is not our intention to define sufficient exposure as we do not have the data to do so.

Now we turn to the relationship between the acoustic measures in the implementation of the contrast and the perception of the native judges. The question that I intend to answer here is if a participant has to implement a reliable contrast on all three acoustic measures in order for the T2 and T3 productions to be perceived as target-like by the native transcribers. From the analysis of the overt tones from Analysis 1, the overt T2 and T3 produced by all participants significantly contrast with each other on the F0 difference, but not all participants made a reliable distinction on the other two measures. Similar results were seen from Analysis 2, where all of the participants credited with having acquired the contrast show a statistically reliable distinction on the F0 difference, but not necessarily the other two acoustic measures. This indicates that a L2 participant does not necessarily have to make a statistically reliable distinction on all three acoustic measures to produce target-like tones. The results from Analysis 1 suggest that a statically reliable contrast of the F0 difference is necessary for the target tones to be overly perceived. And Analysis 2 conforms with this by showing that all of the participants from both elicitations whose T2 and T3 were perceived as target-like made a statistically significant distinction on the F0 difference. As this finding has been attested by two different analytic methods, we suggest that the F0 difference between the contour onset to the turning point is the most crucial acoustic cue to the native transcribers' perception.

In addition to the F0 difference, on the one hand, all participants who have been credited with having the acquired the contrast also showed a reliable distinction on at

least one of the other two acoustic measures. Both the duration and the offset pitch are fairly equally used by the participants as the additional acoustic measure. On the other hand, in the second elicitation of Analysis 1, three participants (KL, HL, and JK, from Table 3.3 in Chapter 3) showed a reliable distinction between the target-like T2 and T3 only on the F0 difference. No contrast was found on the other two measures for those three participants, but the tones were still transcribed as target-like. As a result, we are not certain of the effect of a statistically reliable contrast on an additional acoustic measure in producing a contrast that is transcribed as target-like, but what we can be certain of is that in order to produce target-like T2 and T3 , a L2 participant has to make a distinction on the F0 difference to an extent that is not only statistically significant, but also target-like in the perspective of the native speakers transcription.

6.1.3 Longitudinal Progress

In this section, I will discuss the progress that six participants have made from the first to the second elicitation. This progress or difference between the two elicitations will be presented in terms of two aspects; the percentage of target-like productions will be based on Analysis 1, and the progress of a participant moving from producing no contrast to a covert contrast, or from producing a covert contrast to an overt contrast between T2 and T3 will be examined on the basis of Analysis 2.

First, Analysis 1 shows that out of the six participants who participated in both elicitations, three have demonstrated a statistically higher percentage of target-like T2 and T3 productions in the second elicitation, compared to the initial elicitation. Specifically, statistically more T3 produced by these three participants were identified as

target-like by the native judges in the second elicitation, suggesting that their T3 productions have significantly improved between the two elicitations. In addition, one other participant has also made reliable progress in her target-like T3. Analysis 2 found two participants produced more target-like T3 in the second elicitation, as compared to the first one. The discrepancy regarding the number of participants who have demonstrated a significant improvement between the two analyses is due to the fact of different scoring criteria on the transcription. Analysis 2 adopts a criterion for scoring the transcription, in which a tone was considered target-like when at least one of the native judges transcribed it as target-like, whereas in Analysis 1, a tone was viewed as target-like when both native judges perceived it as target-like. The looser criterion adopted in Analysis 2 naturally results in a higher percentage of target-like productions, thus there is less room for improvement, as compared to Analysis 1.

Overall, the longitudinal analysis suggests that some of the participants have made significant progress in terms of their target-like production; specifically, the improvement mostly arose from the observation that these participants have produced statistically more target-like T3 in the second elicitation, which is eighteen months after the first elicitation.

Next, we move to the discussion of the stages of the participants in the acquisition process. Previous studies (Macken and Barton, 1980; Eckman et al, 2014) have suggested the occurrence of three stages in L1 and L2 acquisition. The three stages are 1) a participant produces no contrast at all, 2) a participant produces a statistically reliable acoustic distinction between the targets, which is yet to be perceived as target-like by the native transcribers, and 3) a participant produces a distinction that

is perceived as target-like by the native transcribers. One of the hypotheses that the current study aims to test is if any of the participants makes longitudinal progress from producing a contrast in one stage to another stage in the two elicitations. That is, we are interested in discussing if any L2 participants progress from not showing any contrast to implementing the T2-T3 contrast covertly, or from implementing the T2-T3 contrast covertly to implementing the contrast overtly. Below, I present a discussion of this progress from the findings of Analysis 2.

Table 6.2: L2 speakers' production, according to transcriptions and acoustic measures

		Percentage of target-like production (80% threshold)			Acoustic measures with statistical results		
Participant	**Elicitation**	**T2**	**T3**	**Transcription**	**Duration**	**Offset pitch**	**F0 Diff.**
WYZ	First	98%	75%	No	-	-	+
	Second	98%	95%	Overt	+	-	+
LKL	First	100%	21%	No	+	-	+
	Second	100%	95%	Overt	-	+	+
FZT	First	100%	54%	No	+	+	+
	Second	98%	85%	Overt	-	+	+
CBR	First	98%	64%	No	+	+	+
	Second	100%	80%	Overt	+	+	+
NJ	First	98%	80%	Overt	+	-	+
	Second	90%	95%	Overt	+	+	+
MXS	First	98%	80%	Overt	+	+	+
	Second	100%	95%	Overt	+	+	+

Table 6.2 above reports the results of the six participants who participated in both elicitations with the percentage of target-like T2 and T3 and the acoustic measures in which a statistically reliable distinction was found. The first column lists the participants, followed by the elicitation in the second column. The third column presents the percentage of target-like productions and whether a particular participant was credited with having acquired the contrast based on the 80% target-like threshold for scoring the transcription. The last three columns show whether the acoustic analysis reveals a statistically reliable distinction on the acoustic measures.

From the table, it is seen that the first four participants were not credited with having acquired the contrast in the first elicitation because they failed to produce at least 80% of target-like T3. But the acoustic analysis indicates that they all have made a statistically reliable distinction on at least one of the measures. Therefore, in the first elicitation, those four participants were in the intermediate stage proposed by the previous studies, where they did not produce the target-like contrast, but produced a covert contrast between T2 and T3. In the second elicitation, all four of them produced at least 80% target-like T3; thus, they have been given credit for having acquired the contrast. In this case, according to the assumption that a participant is considered producing an overt T2-T3 contrast when he/she is credited with having acquired the contrast, it is suggested that they have made progress from implementing the T2-T3 contrast covertly to implementing it overtly based on the two elicitations. Regarding the two participants who have presented an overt contrast between T2 and T3 in the first elicitation, their target-like T2 and T3 both remained more than 80% of all the targets

produced in the second elicitation. To briefly conclude, all six participants who have participated in the longitudinal analysis have been credited with having acquired an overt contrast between T2 and T3 at the end of the second elicitation.

Let us proceed to discuss the longitudinal changes in the implementation of the contrast on the acoustic measures. Four of the participants have shown different patterns of the acoustic implementation of the contrast between the two elicitations while the other two have implemented the contrast significantly on all three acoustic measures in both elicitations. Two (WYZ and NJ) of the four participants with different implementation patterns have made a covert contrast on one more acoustic measure in the second elicitation than in the first one. WYZ initially only shown a reliable distinction on the F0 difference, but have also shown such a distinction on the duration measure in the second elicitation as his target-like T3 exceeded the 80% threshold. NJ, who was considered to have acquired the contrast in the first elicitation but only showed a reliable acoustic contrast on two of the measures, implemented the contrast statistically on all three measures. The other two participants (LKL and FTZ) with different patterns of the acoustic implementation between the elicitations did not shown any additional acoustic measure in making a reliable contrast. Instead, LKL showed such a contrast on two acoustic measures in both of the elicitations. In addition to the F0 difference, which remains as a contrastive measure in both elicitations, she also implemented the contrast on the additional duration measure in the first elicitation, but only on the additional offset pitch in the second elicitation. FZT showed a covert contrast on all three measures in the first elicitation, but only implemented it on two measures, the

offset pitch height and the F0 difference in the second one in which she was credited with having acquired the contrast.

The finding that the six participants demonstrated various patterns of change in the systematic implementation of the contrast in terms of the acoustic measures shows that there is not a uniform trajectory in terms of which acoustic measure will be implemented before or after another. Yet, this finding demonstrates the nature of L2 participants' IL: systematicity and variability. The tone productions from all six L2 participants are systematic because every participant shows a statistically reliable distinction on at least one measure in the first elicitation and on at least two measures eighteen months after the initial elicitation, which always includes a contrast in the F0 difference. They are individually variable in that some participants implemented the contrast on different combinations of the measures and they changes the patterns of the acoustic implementation between the elicitations when their target-like productions have improved longitudinally. It supports that IL of every L2 participant must be viewed independently from one another, in which is the only way that its variability can be analyzed.

6.1.4 Addressing the Hypotheses from the production task

To conclude the discussion of the production findings, we address the two hypotheses proposed in Chapter 2, which are repeated here for convenience as (3) and (4):

(3) L2 learners of Mandarin may implement a statistically reliable acoustic distinction between T2 and T3 that is not perceived by native speakers of Mandarin.

(4) Some L2 participants will progress from not showing any contrast to implementing the T2-T3 contrast covertly, but not vice versa; or from implementing the T2-T3 contrast covertly to implementing the contrast overtly, but not vice versa.

As discussed in the previous section, the findings we gathered from the production task support both of the above hypotheses. Hypothesis (3) is strongly supported in that two analyses show that some participants in both elicitations have implemented the T2-T3 contrast acoustically to an extent that is statistically reliable when the contrast is not perceived as target-like by the native transcribers. Hypothesis (4) is also supported by the finding that all four participants who participated in both elicitations and were not credited with an overt T2-T3 contrast initially have developed a longitudinally from implementing the T2-T3 contrast covertly to implementing the contrast overtly. As in the first elicitation there was no participant who exhibits no contrast at all, we can not draw any conclusion on whether a participant can progress from making no contrast at all between T2 and T3 to implementing the contrast covertly. However, it is plausible to argue the occurrence of such a progress because the occurrence of an intermediate stage of covert contrast is attested in the current study. One can logically assume that a L2 learner who at the very beginning possesses no knowledge of pitch movements or tone would be able to progress to making the T2-T3

contrast covertly, if given sufficient exposure. Nevertheless, it essentially depends on the IL of a L2 learner and how the IL develops. In the next section, the findings of the identification tasks will be discussed.

6.2 Identification Task

The goal of the identification task was to investigate whether the L2 participants who produced an overt contrast identified T2 and T3 statistically better than the L2 participants who produced a covert contrast. The findings indicate no significant difference in the percentage of correct identification of the targets between the L2 overt group and the L2 covert group, suggesting that the participants credited with an overt contrast did not identify T2 and T3 more accurately than the participants who implemented a covert contrast. With this finding, the hypothesis proposed for the identification task, stated below in (5), is falsified.

(5) The L2 participants who produce an overt contrast between T2 and T3 will identify the T2 and T3 statistically better than the L2 participants who implement the contrast covertly.

In other words, while the two groups of participants were at different stages in terms of the production of T2 and T3, it is shown that their perception is not statistically different. While the covert groups' production lags behind the overt group, their perception accuracy of T2 and T3 does not significantly fall behind. This lack of any significant difference between the overt and the covert participant can be accounted for from the perspective of the relationship between the perception and the production of a L2 contrast. Some studies (Sheldon and Strange, 1982; Bradlow et al, 1997) have

suggested that in the relationship between L2 production and perception, one aspect will precede the other aspect. Since we found that L2 participants at different production stages demonstrated a difference that is not statistically reliable in identifying T2 and T3, what this might suggest is that the perception and the production of the T2-T3 contrast may develop in different fashions. The correct T2 and T3 identification of both groups is at least 80% when two options were given to the participants, which indicates that the covert participants have identified T2 and T3 at a similar accuracy level as the overt participants. In other words, the covert participants can sufficiently and correctly perceive T2 and T3, but their production, especially T3 production fails to exhibit an overt contrast. One plausible explanation is that for the covert participants, the perception of T2 and T3 may develop faster than the production. The finding is consistent with the results of Bohn and Flege (1997), where the authors found that when compared to the experienced L2 learners, the inexperienced L2 learners sufficiently distinguished the English /æ/ and /ɛ/ while they can not systematically produce the contrast.

In terms of the discrepancy between correctly identifying the two tones, the finding aligns with the production results. It is not the intent of the current study to assert a correlation between the perception and production of the T2-T3 contrast, but a clear pattern can be seen here. From the production of most of the participants, more T2 were perceived as target-like than T3 by the native transcribers, which indicates it is easier to produce T2 than T3 in a target-like fashion. From the identification task, when the participants were given four tone options to choose among, T2 targets were correctly identified statistically better than T3 targets for both groups. This suggests that

as the participants have more difficulty producing target-like T3 than T2, they also tend to have more difficulty in correctly identifying T3 than T2. What this pattern might be suggesting is that how accurately L2 participants identified T2 and T3 and how target-like the participants' production were perceived may be related, but we do not have sufficient data and analysis to further support the observation.

On the other hand, when the participants were given only two options, T2 and T3, to choose from, the correct identification of T3 increased significantly to a percentage that is similar to the correct identification of T2, which remained similar in both option conditions. That is, when only provided with T2 and T3 to choose from, both groups successfully identified T3 at a significantly higher percentage than in the four-option condition. Specifically, the overt group reached 98% of the correct identification and the covert group correctly identified approximately 75% of the T3. Since the two groups did not choose T2 to label the T3 targets, respectively at 98% and 75% of the attempts, it suggests that they can differentiate T3 from the T2. This indicates that discriminating the two target tones is easier than identifying the target tones. This finding can be crucial because it may suggest that there is another tone other than T2 and T3 that the L2 participants choose to label as the T3 targets in the four-option condition. For eleven of the twelve participants, over 80% of the incorrect responses in the T3 identification were identified as T4. Among the eleven, eight labeled over 90% of the incorrectly identified T3 as T4. From the misidentification of T3, we can see that instead of T2, T4 is the tone that was most frequently confused with T3 in the perception of most participants in this study. Without detailed analysis of the acoustic similarity of T3 and T4, no conclusion can be drawn to account for this finding. However, this unexpected

finding can perhaps be accounted for by the acoustic similarity of the partial contour of T3 and T4. T3 is a contour tone that falls in the first half of the contour before hitting the turning point and rising at the end, and T4 is a falling tone that starts high and falls through the entire contour. Although T3 and T4 start at different pitch heights, both of them fall in the contour, partially or completely. This factor may attribute to the confusion between T3 and T4. Also, other studies (Li et al, 2006; Sanders, 2008) have argued that in some dialects of Mandarin, including Taiwanese Mandarin, T3 is in fact a low falling tone, in which the rising part of the contour in the Beijing dialect is not implemented. The speaker of the stimuli in the perception tasks is a native speaker of Taiwanese Mandarin. Therefore, this is assumed to be a factor that affects the perception of the L2 participants, which led to the confusion of T3 and T4 by some of them. Nevertheless, without further investigation of the relationship between L2 perception and acoustic variations of T2, T3 and T4, we cannot suggest any additional reason for the confusion between T3 and T4.

The next section discusses the findings of the discrimination task that tested the participants' perceptual sensitivity to subtle F0 distinction on the continuum of the T2 and T3 contrast.

6.3 Discrimination Task

This section presents a discussion of the key findings of the discrimination task and bears on the hypothesis proposed to investigate the participants' perceptual sensitivity to fine-grained acoustic differences. First, we found a main effect of the base tone. All four participant groups showed very similar results in terms of how the base

tone affects their perceptual discrimination. The participants discriminated the T2-based stimuli statistically better than the T3 based stimuli. When two T3-based stimuli that are three intervals apart were paired, the participants always perceived them as different tones in less than 50% of the attempts, while two T2-based stimuli which are three intervals apart were perceived as different tones at 90% or higher in the attempts. The discrepancy of the responses to the two base tones suggests that there is some auditory information in T2 or T3 as a base tone that affects the participants judgment on the stimuli. Specifically, one possible reason that can contribute to the discrepancy is that some acoustic feature of T3 that remains in the stimuli after the pitch manipulation that hindered the participants' ability to perceive the F0 difference in of the stimuli. Mandarin T3 is widely realized as a laryngealized low tone with a creaky voice quality, and many studies have agreed that the creaky quality is a useful perceptual cue in tonal recognition (Davidson, 1991; Yu, 2010). As the stimuli used in the discrimination task were only modified on the F0 contour, but not on the voice quality, it is likely that such a laryngealized feature of T3 remained in the T3-based stimuli after pitch manipulation.

Also, the discrimination task revealed an interaction between the interval and the speaker groups. A significant distinction in the percent of correct distinction was found between stimuli one interval apart and stimuli three intervals apart, for the NS Mandarin group and the L2 overt group. That is, the NS Mandarin speakers and the L2 overt participants are similar to each other in that when two stimuli are three intervals apart, they were statistically more likely to perceive the stimuli as different than when the stimuli distance was one or two intervals. However, such an effect was not present in the L2 covert group and the non-tonal speaker group. The stimuli that are three intervals

apart essentially have the identical F0 values of a real T2 and T3, which suggests that only the NS and L2 overt group can reliably discriminate T2 and T3 when the two tones were presented in parallel. The L2 covert group and the non-tonal speaker group failed to reliably distinguish stimuli that resemble an actual T2 and T3. Since no significant main effect of the speaker groups was found, it may seem that the hypothesis that the performance of L2 participants who implement a covert contrast will be systematically different from that of the native speakers of Mandarin is falsified. However, in spite of the lack of statistically significant difference among the groups, the L2 covert group and the NS speaker group in fact showed distinct patterns in the difference between the discriminating stimuli that are three intervals apart and one interval apart. Moreover, the L2 overt participants resembled the NS Mandarin group more than the L2 covert group did in discriminating stimuli that are three intervals apart versus one interval apart. Considering everything together, the L2 overt participants' performance differs from the L2 covert participants' in terms of how similar their results are to the NS speakers.

Another key finding is that, from the native speakers' results, a significant effect of the intervals is only found between stimuli with one-interval distance and stimuli with a three-interval distance; such an effect is not found to be significant between stimuli with one-interval distance and two-interval distance, nor between stimuli that are two intervals apart and three intervals apart. As mentioned in the previous paragraph, stimuli with a three-interval distance are essentially T2 and T3. Therefore, it was expected that the native speakers would show significantly different results between discriminating T2 and T3 and discriminating stimuli that are one interval apart. Therefore, no clear implication can be drawn concerning where the perceptual boundary of T2 and

T3 is in terms of the number of intervals. Nevertheless, if we return to Figure 5.3 in Chapter 5 (repeated below for the convenience), we see that the percentage of perceiving the T2-based stimuli as different. What was found was that NS perceived two stimuli that are only one interval apart as different at a below chance level(41%), but the chances of perceiving stimuli with two-interval distance as different tones dramatically rises to above 80%. This may imply that while two stimuli that are one interval apart may still be considered in the same tone category, stimuli that are two intervals apart are very likely to fall in two different categories. This contradicts the finding from the statistical test on the effect of intervals within the NS Mandarin group in the previous paragraph, that only the difference between stimuli with one-interval distance and stimuli with three-interval distance is statistically reliable. The contradiction is because it accounts for the different intervals within NS Mandarin group regardless of the baseline tone. As we see, T2-based and T3-based stimuli showed very different results. I have conducted a one-way ANOVA model for the NS Mandarin group's response to T2-base stimuli only. The statistical result is presented in Table 6.3.

Figure 5.3 Percent of correct responses to T2-based paired stimuli with 1, 2 or 3 intervals apart by the 4 speaker groups

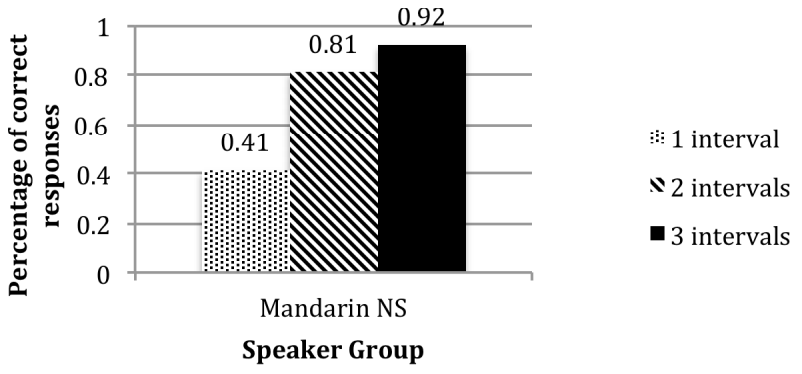


Table 6.3: Results of one-way ANOVA of the percentage of response "different" of T2-based stimuli from the NS Mandarin group

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Interval	2	3.468	1.734	86.51	<2e-16 ***
Residuals	69	1.383	0.020		

---

	diff	lwr	upr	p adj
Interval_2-Interval_1	0.3997500	0.30184693	0.4976531	0.0000000
Interval_3-Interval_1	0.5112083	0.41330527	0.6091114	0.0000000
Interval_3-Interval_2	0.1114583	0.01355527	0.2093614	0.0218037

Signif. codes: 0 '***'

The results have indicated that when focusing on the T2-based stimuli, the distinction between any of the two intervals is significant for the NS Mandarin group. However, the distinction between the stimuli with one-interval distance and the stimuli with a two-interval distance, and the distinction between the one-interval distance and the three-interval distance is statistically more significant than the difference between the three-interval distance and the two-interval distance. This suggests that, compared to stimuli with one-interval distance, it is more probable that the NS Mandarin participants may have perceived stimuli that are two or more intervals apart as tones from two tone categories, instead of within the same category.

Now I proceed to address the two hypotheses proposed concerning the discrimination task, as repeated below.

(6) The performance on the discrimination task by L2 participants who produce an overt contrast between T2 and T3 on the production task will not be reliably different from the performance on the discrimination task of the native speakers of Mandarin.

(7) The performance of L2 participants who implement a covert contrast will be systematically different from that of the native speakers of Mandarin.

The hypothesis stated in (6) is supported by the finding that no significant effect of the speaker groups was found. The performance on the discrimination task by L2 participants who are credited with having acquired the contrast is not reliably different from the results by the native speakers of Mandarin. Also, as discussed above, the L2 overt participants are similar to the native speaker of Mandarin, in that when two stimuli are three intervals apart, they were statistically more likely to perceive the stimuli as different than when the stimuli distance is one or two intervals.

The hypothesis stated in (7), on the one hand, was not supported because there was a lack of any reliable distinction in the performance among the speaker groups. On the other hand, the L2 covert participants and the NS speaker group in fact showed distinct patterns in discriminating between stimuli that are three intervals apart and one interval apart. Although this finding cannot directly support the hypothesis, we do find that the native speakers of Mandarin discriminated stimuli with one-interval distance and stimuli with three-interval distance at a statistically different rate whereas the L2 participants who produced a covert contrast did not.

6.4 Implications

This section presents the implications of the findings from the current study. I focus on four issues pertaining to the data on L2 acquisition of Mandarin T2 and T3. They are 1) the relationship between L2 production and the perception of the TL native speakers, 2) the stages in the process of acquiring a L2 tonal contrast, 3) the

relationship between L2 production and perception, and 4) the implications for L2 instruction.

6.4.1 L2 production and the perception of the TL native speakers

One of the commonly used methods to determine whether a L2 learner has acquired the production of an L2 phonemic contrast is to see whether the contrast has been perceived by native speakers of the target language (Flege, 1987; Eckman & Iverson, 1994). In other words, L2 acquisition has sometimes largely depended on the perception of the TL native speakers. The findings of the current study indicate that some L2 participants have demonstrated a statistically reliable distinction between the Mandarin T2 and T3 when the contrast was not transcribed as target-like by the Mandarin native speaking transcribers. By showing a covert contrast implemented by some of the L2 participants via two different analyses, the findings align with and strengthen what has been found in several other studies on L2 covert contrast conducted in recent years, that L2 participants may be producing a contrast that is not perceived as target-like. When the commonly-used method, which considers L2 production on the basis of TL native speakers' perception, is employed, the interlanguage of a L2 participant is essentially being viewed from the target language and the independency of the interlanguage is fundamentally overlooked.

The empirical evidence of a covert contrast shows the discrepancy between what is being perceived from the perspective of the target language and what actually occurs the interlanguage system of a L2 participant. Because of the discrepancy, whether a L2 learner can produce an L2 contrast solely depends on whether the native speakers of

the target language perceive it as target-like. We may be running the risk of missing some aspects in the development of a L2 learner's process of acquiring an L2 contrast. Thus, in order for L2 interlanguage to be viewed objectively, the perception of the TL native speakers should not be the only source of determining whether a L2 learner can produce an L2 phonemic contrast. Along with the perception of the TL native speakers, the acoustic measurements should also be taken into account. With the three acoustic measures in the current study, the findings suggest the native speakers of the TL language are most sensitive to the F0 difference between the contour onset and the turning point in identifying T2 and T3. As T3 is categorized as a low tone with a creaky voice quality, a substantial F0 drop in the tone contour serves as an obvious auditory cue of a target-like T3 to native speakers' perception. However, as reported in the production experiment, some L2 learners have implemented a statistically reliable contrast on one or both of the other two acoustic measures that do not seem to be as crucial as the F0 difference in native speakers' perception. In such instances, the contrast that has been established in a L2 learner's interlanguage is not properly recognized, which may lead to misunderstanding of the contrast by a native speaker and misinterpretation of the development of acquiring the L2 contrast by a linguist.

6.4.2 The stages in acquiring a L2 contrast

The current study provides empirical evidence of the occurrence of a covert contrast. With the three stages initially identified by Macken and Barton (1980), if the findings can be interpreted in a similar fashion as the covert contrast attested in the field of L1 acquisition and disordered speech, the occurrence of a covert contrast offers an

empirical example of an intermediate stage in the process of acquiring a L2, in this case, tonal contrast. The presence of an intermediate stage implies that L2 learners may start from a stage where no tonal contrast is established in their interlanguage, and then reach an intermediate stage of implementing a covert tonal contrast before systematically producing an overt contrast. While the current study did not obtain any data from L2 participants with no contrast at all in the first production elicitation, it cannot evidence the progress from an initial acquisition stage to an intermediate stage, the findings of the longitudinal analysis empirically demonstrate that some L2 participants have progressed from implementing a covert contrast to being credited with producing an overt contrast. The acquisition of an L2 has always been considered to be an ongoing progress that is constantly developing because it requires time and exposure to the L2 for its learners to acquire how to implement a L2 phonemic contrast. This study sheds lights on the long-standing idea and the possibility for L2 learners to achieve the final stage of producing an L2 contrast overtly. In other words, an intermediate stage of implementing a covert contrast points to the fact that a contrast has been established in the interlanguage and can be an indication that a learner's interlanguage is moving towards establishing a contrast that is target-like in the L2.

6.4.3 Relationship between the production and perception

The results of the identification task indicate that regardless of whether the L2 participants have been credited with having acquired the T2-T3 contrast, they performed similarly in identifying the targets. The L2 covert participants' ability to correctly identify T2 and T3 was at the same level as the L2 overt group since no

statistically reliable distinction was found between the two groups, whereas the production was at different levels in terms of whether they have been credited with producing an overt contrast. Similarly, no statistically reliable distinction was found between the two L2 groups on the discrimination task. The proposal I am offering here to account for the discrepancy between the production and the perception performances from the two groups of L2 participants is that the two aspects, production and perception, of acquiring T2-T3 in Mandarin develops in different fashions. One interpretation to further explain the discrepancy is that the L2 overt participants' perception did not match up with their own production. In the production task, a participant was credited with producing an overt contrast when at least 80% of both T2 and T3 were perceived as target-like, but in the identification task with four options available to choose from, the overt participants have only correctly identified less than 40% of the T3. From this perspective, for the L2 overt participants, their perception has fallen behind their own production of T3. Regarding the L2 covert participants, their production and identification of T3 both did not reach the 80% target-like criterion. Therefore, we cannot assume one aspect leads or falls behind the other aspect for the covert participants.

One intriguing finding worth pointing out is that the discrepancy between acquiring T2 and T3 is parallel in production and identification. For both the overt and covert participants, T3 almost always is harder to produce in a target-like fashion than T2, and T3 is also more difficult to correctly identify than T2.

In short, although the two L2 groups differ in whether they implemented the T2-T3 contrast covertly or overtly, their ability to identify T2 and T3 and their sensitivity to

the subtle difference on the F0 continuum of T2 and T3 do not systematically differ from each other.

6.4.4 Implications on L2 instruction

I now provide two pedagogical implications that can be applied in L2 instruction to assist the acquisition of Mandarin T2 and T3. The first and the most important implication is that when any L2 production of T2 and T3 fails to be perceived as target-like by a native speaker of Mandarin, this does not necessarily indicate that a L2 learner has not established a contrast between T2 and T3. It is possible that this learner is in the intermediate stage of the acquisition process. L2 learners and instructors can direct more attention to the progressive and ongoing nature of the learning of a L2 contrast. In L2 instruction, while it is mostly agreed that L2 acquisition is a constantly developing process, learners and instructors generally expect the acquisition of a L2 phonemic contrast to be an off-and-on sequence, that a learner would either produce a target-like contrast or fail to produce a contrast at all. Under such an expectation, L2 learners are often unaware of the progress that they are making themselves; thus, can become discouraged during the process. The demonstration of an intermediate stage may serve as a reminder for L2 learners and instructors that such an expectation should be less imposed in the acquiring an L2 phonemic contrast. If more attention and awareness is directed to the fact that they might be implementing the contrast covertly and the possibility of being able to improve from an intermediate stage to one where the contrast is overtly considered as target-like, L2 learners can spend more effort on the

specific aspects of the tone implementation in order to make progress in this process and ultimately achieve target-like productions, which is the intention of L2 acquisition.

Secondly, the findings strongly indicate that a substantial F0 drop in the beginning of the tone contour is the major acoustic cue to produce a target-like T3 and disambiguate T2 and T3. This is the specific acoustic aspect that L2 learners and instructors of Mandarin can focus on in the instruction. Our results have also shown that some participants who were credited with having acquired an overt contrast have not demonstrated a statistically reliable distinction in the offset pitch height. Therefore, I suggest that in L2 instruction, in order to solve the difficulty and the confusion of producing T2 and T3, it is not necessary to introduce the later half of the contour in T3, which is categorized as a rising movement of pitch in the Beijing Mandarin dialect. T3 can be simply categorized as a low falling tone, as that is how it is implemented in some of the dialects of Mandarin. In this case, T2 and T3 would be less confusable in terms of their contour shape and the effort can be focused on maintaining a low pitch in producing T3 and on rising the pitch for T2. By suggesting this, the contrastive cue of offset pitch height between T2 and T3 can still be upheld because the offset pitch of T3, without the rising end, would likely remain lower than that of T2, whose major contour lies on the rising movement of pitch.

Chapter 7

Conclusion

This chapter concludes the current study on the acquisition of Mandarin T2 and T3 by L2 Mandarin learners whose native language is American English. Three experiments were designed and conducted primarily to investigate the following three topics: 1) the occurrence of a covert contrast in the process of acquiring the T2-T3 contrast in L2 Mandarin acquisition, 2) the longitudinal development of L2 productions of T2 and T3 in terms of the stages in the acquisition process by a L2 learner, and lastly, 3) whether the perception of the L2 participants was similar to their L2 production on the basis of whether they have produced an overt or a covert contrast between Mandarin T2 and T3. The primary goal is to understand the interlanguage system of each L2 learner from the perspective of its being independent from the native language. Specifically, a covert contrast refers to a statistically reliable distinction between two L2 phonemic targets produced by a L2 learner, which is not perceived as target-like by native speakers of the target language. The attestation of a covert contrast demonstrates that, in some cases, it is important to evaluate L2 production independently from the native speakers' perception of the target language, and that L2 acquisition is an ongoing process in which an intermediate stage occurs before fully acquiring the target contrast.

Chapters 3, 4 and 5 each reports the details of the three experiments in the study. In Chapter 3, two identical production tasks were conducted at two different points in time for the longitudinal analysis of the data. All production data were analyzed by measuring three acoustic cues that have been shown to systematically differentiate

Mandarin T2 and T3; they are the F0 difference between the contour onset and the turning point, the duration from the contour onset to the turning point, and the offset pitch height. Two analyses of the production data, differing in the scoring criterion and the inclusion of the 80% criterion for target-like native transcription, both suggest that some participants implemented a covert contrast between T2 and T3 on at least one of the acoustic measures. Among the three acoustic measures, the F0 difference appears to be the most crucial acoustic cue that can be attributed to target-like transcriptions by the native speakers of Mandarin. In addition, the longitudinal analysis supports the proposed hypothesis, showing that in a timespan of seventeen months, the L2 participants in this study have made progress from implementing a covert contrast to producing an overt contrast that has been perceived as target-like by the native speakers of Mandarin on more than 80% of the attempts. Another finding that aligns with many studies on Mandarin T2 and T3 is that it is more difficult for the participants to produce target-like T3 than target-like T2.

The findings of the two perception tasks do not directly show any similar discrepancy found in the production task between the L2 participants who produced a covert contrast and those who produced an overt contrast. Chapter 4 reports the details of an identification task that asked the L2 participants to identify T2 and T3 targets in two different conditions, one in which four tone options were provided and one with only two options, T2 and T3. The finding of the identification task shows no significant distinction between the two L2 participant groups, indicating that while the two groups show distinct patterns in term of whether they produced an overt contrast or a covert contrast, their perceptual performance in identifying T2 and T3 does not statistically

differ from each other. What was found in the identification task that is similar to the production task is that T2 are statistically more often correctly identified than T3 in both of the option conditions.

Chapter 5 focuses on the perceptual sensitivity to subtle acoustic differences on the F0 continuum of T2 and T3 via a discrimination task in which four groups of participants, the two L2 participant groups, a group of native speakers of Mandarin, and a group of speakers of a non-tonal language, discriminated two targets of 0-3 interval distances on the F0 continuum. Specifically, the primary aim of the discrimination task was to investigate whether the L2 overt participants' perceptual sensibility to fine-grained F0 difference resembled that of the native speakers of Mandarin more than that of the L2 covert participants. The finding shows no main effect of the speaker groups, but suggests that the L2 overt participants resemble the Mandarin native speakers, who have shown a different result from the L2 covert participants in the statistical discrepancy of the discriminating performance between the stimuli that are one interval apart and stimuli that are three intervals apart. Since the task did not find a reliable difference between the L2 covert participants and the native speakers of Mandarin, my hypothesis that the performance of L2 participants who implement a covert contrast will be systematically different from that of the native speakers of Mandarin is not strongly supported, but can still be upheld as the two groups show different results in discriminating the stimuli that are one interval apart and those that are three intervals apart. It is important and necessary to continue the investigation of how L2 participants at different levels interact with each other, or the native speakers of Mandarin in terms of their perceptual sensibility to fine-grained acoustic differences. More data need to be

gathered in order to decide the specific way in which the speaker groups interact with the intervals.

Another point worth clarifying is that the current study did not aim to investigate whether the production performance of the two participant groups correlates with their respective perception data in a statistical fashion. What has been done is that the L2 perception data were analyzed based on the groups categorized by the L2 participants' production performance, i.e. whether they produced a covert or an overt contrast. It would also be interesting to analyze how L2 perception and production associate with each other in a statistical correlation model from the perspective of whether a L2 participant is producing a covert or an overt contrast between Mandarin T2 and T3.

Finally, I would like to conclude the dissertation by highlighting a few pedagogical implications of the study. First, I want to reiterate the finding that L2 participants can produce a covert contrast when a target contrast is not perceived as target-like by the native speakers of Mandarin. Instead of completely focusing on whether a L2 learner produces a target-like tone, some attention should be brought to the intermediate stage of tone acquisition in L2 instruction of Mandarin, and L2 learners should be given some degree of credit for making a covert contrast. Secondly, it is suggested that L2 learners can improve from implementing a covert contrast to producing an overt contrast after a certain period of time and exposure. It should be the aim of future studies to investigate how much exposure can lead to such a progress. But what is certain is that L2 learners should be aware of the progress that they are making or can make in acquiring T2 and T3. Lastly, L2 production and perception of T2 and T3 may develop in different fashions due to the lack of a reliable distinction between the L2 participant groups at two different

levels in identifying T2 and T3 and perceiving the subtle acoustic distinction on the F0 continuum of T2 and T3.

With the occurrence of an intermediate stage in the process of acquiring Mandarin T2 and T3, I conclude that L2 acquisition of a tonal contrast is not a categorical process, but more similar to a gradient one.

References

Best, C. T. (1995). A Direct Realist View of Cross-Language Speech Perception. *Speech perception and linguistic experience: Issues in cross-language research*, 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*, 13-34.

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

Blicher, D. L., Diehl, R. L., & Cohen, L. B. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics*.

Bohn, O. S., & Flege, J. E. (1997). Perception and production of a new vowel category by adult second language learners. *Second-language speech: Structure and process*, 53-73.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English/r/and/l: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*(4), 2299-2310.

Davison, D. S. (1991). An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA Working Papers in Phonetics*, *78*, 50-57.

Eckman, F., & Iverson, G. (1994). Pronunciation difficulties in ESL: Coda consonants in English interlanguage. In M. Yavas (Ed.), *First and Second Language Phonology,* San Diego: Singular Publishing Group.

Eckman, F., Iverson, G., & Song, J. (2014). The role of covert contrast in the acquisition of L2 phonemic contrasts. *International Review of Applied Linguistics in Language Teaching*.

Elliott, C. E. (1991). The relationship between the perception and production of Mandarin tones: An exploratory study. *University of Hawai'i Working Papers in English as a Second Language 10 (2)*.

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

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 233-277.

Flege, J. E., & Schmidt, A. M. (1995). Native speakers of Spanish show rate-dependent processing of English stop consonants. *Phonetica*, *52*(2), 90-111.

Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America*, *106*(5), 2973-2987.

Goldinger, S., Pisoni, D., & Luce, P. (1996). Speech perception and spoken word recognition: Research and theory. In N. Lass (Ed.), *Principles of Experimental Phonetics,* St. Louis: Mosby.

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

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099-3111.

Howie, J. M. (1976). *Acoustical studies of Mandarin vowels and tones* (No. 6). Cambridge University Press.

Ingram, J. C., & Park, S. G. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of phonetics*,*25*(3), 343-370.

Jongman, A., Wang, Y., Moore, C. B., & Sereno, J. A. "Perception and Production of Mandarin Chinese Tones." *Handbook of Chinese Psycholinguistics*. E. Bates, L.H. Tan, and O.J.L Tzeng (eds.). Cambridge University Press. 2006.

Kluge, D. C., Rauber, A. S., Reis, M. S., & Bion, R. A. H. (2007). The relationship between the perception and production of English nasal codas by Brazilian learners of English. In *Interspeech* (pp. 2297-2300).

Li, A., Xiong, Z., & Wang, X. (2006). Contrastive study on tonal patterns between accented and standard Chinese. In *Proceedings of the 5th International Symposium on Chinese Spoken Language 2006*.

Li, F. Edwards, J., & Beckman, M. (2009). Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. *Journal of Phonetics*, *37*, 111-124. doi: Elsevier

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54*(5), 358-368.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.

Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th international congress of phonetic sciences* (Vol. 563, pp. 563-567). Academia Prague.

Macken, M. A., & Barton, D. (1980). A longitudinal study of the acquisition of the voicing contrast in american-english word-initial stops, as measured by voice onset time. *Journal of Child Language*, *7*, 41-74.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33-B42.

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

Ohala, John J. "Phonetic explanations for nasal sound patterns." *Nasálfest: Papers from a symposium on nasals and nasalization*. 1975.

Peperkamp, S. & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. In *INTERSPEECH* (pp. 161-164).

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

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*(2), 285-290.

Sanders, R. M. (2008). Tonetic sound change in Taiwan mandarin: The case of tone 2 and tone 3 citation contours. In *20th North American Conference on Chinese Linguistics*.

Sato, C. J. (1984). Phonological processes in second language acquisition: Another look at interlanguage syllable structure. *Language Learning*, *34*(4), 43-58.

Scobbie, J. E., Gibbon, F., & William, J. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. *Papers in laboratory phonology V: Acquisition and the lexicon*, *5*, 194.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(03), 243-261.

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

Shen, X-N. S. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics,* 18, 281-295.

Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3.*Language and speech*, *34*(2), 145-156.

Shen, J., Deutsch, D., & Rayner, K. (2013). On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements. *The Journal of the Acoustical Society of America*, *133*, 3016.

Shih, C., Lu, H.-Y., Sun, L., Huang, J.-T., & Packard, J. (2010). An Adaptive Training Program for Tone Acquisition. In *Proceedings of Speech Prosody 2010*.

So, C. K. L. (2006). *Effects of L1 prosodic background and AV training on learning Mandarin tones by speakers of Cantonese, Japanese, and English*(Doctoral dissertation, Department of Linguistics-Simon Fraser University).

Tarone, E. (1978). The phonology of interlanguage. *Understanding second and foreign language learning*, 15-33.

Wang, Y. & Jongman, A. (2006). L2 Acquisition and Processing of Mandarin Tone. *The Handbook of East Asian Psycholinguistics: Volume 1, Chinese*. Li, P., Tan, L., Bates, E. and O.J.L Tzeng (eds.) Cambridge University Press.

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

Yu, K. M. (2010). Laryngealization and features for Chinese tonal recognition. In Proceedings of Interspeech 2010.

Zhang, Y. (2011). *Distinction between mandarin tones 2 and 3 for L1 and L2 listeners*. In Zhuo Jing-Schmidt (Ed.), *Proceedings of the 23rd North American Conference on Chinese Linguistics* (pp. 84-96).

Appendices

Appendix A: Reports of Individual Participant's acoustic measurements in Elicitation 1 Analysis 1

Participant DSY: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean	Significance test
F0 Difference (Hz)	NT2	39.4	t=0.115, p=0.9, p>.05
	NT3	41.4	
	Overt T2	20.9	t=2.362, p=0.024, p<.05
	Overt T3	39.4	
	All T2	24.7	t=2.08, p=0.04, p<.05
	All T3	41.1	
Duration to the Turning Point	NT2	0.23742	t=2.302, p=0.02, p<.05
	NT3	0.32778	
	Overt T2	0.19913	t=2.764, p=0.009, p<.05
	Overt T3	0.35564	
	All T2	0.20699	t=5.1, p=0, p<.05
	All T3	0.33135	
Offset Pitch(Hz)	NT2	291.05	t=-1.545, p=0.13, p>.05
	NT3	275.55	
	Overt T2	306.98	t=-3.3, p=0.002, p<.05
	Overt T3	262.87	
	All T2	303.71	t=-4.9, p=0, p<.05
	All T3	273.92	

Participant BHR: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	--	N/A
	NT3	--	
	Overt T2	25.405	t=4.534, p=0, p<.05
	Overt T3	69.618	
	All T2	25.4	t=1.123, p=0.265, p>.05
	All T3	29.4	
Duration to the Turning Point	NT2	--	N/A
	NT3	--	
	Overt T2	0.27413	t=4.076, p=0, p<.05
	Overt T3	0.50391	
	All T2	0.27413	t=3.628, p=0, p<.05
	All T3	0.35709	
Offset Pitch	NT2	--	N/A
	NT3	--	
	Overt T2	164.3	t=-2.263, p=0,.0293 p<.05
	Overt T3	140.59	

	All T2	164.3	t=-2.086, p=0.0404, p<.05
	All T3	158	

Participant XLX: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	40.54	t=-0.009, p=0.002, p>.05
	NT3	40.38	
	Overt T2	12.11	t=-3.806, p=0, p<.05
	Overt T3	45.19	
	All T2	20.42	t=2.181, p=0.0322, p<.05
	All T3	41.17	
Duration to the Turning Point	NT2	0.425	t=-0.069, p=0.945 p>.05
	NT3	0.421	
	Overt T2	0.287	t=13.5, p=0, p<.05
	Overt T3	0.482	
	All T2	0.326	t=3.762, p=0, p<.05
	All T3	0.432	
Offset Pitch	NT2	122.04	t=1.722, p=0.0924, p>.05
	NT3	130.51	
	Overt T2	134.7	t=2.748, p=0.0097, p<.05
	Overt T3	165.2	
	All T2	131.13	t=0.968, p=0.336, p>.05
	All T3	135.85	

Participant WZW: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	11.194	t=0.693, p=0.492, p>.05
	NT3	16.681	
	Overt T2	8.911	t=2.983, p=0.005, p<.05
	Overt T3	28.5	
	All T2	9.263	t=2.43, p=0.0174, p<.05
	All T3	17.683	
Duration to the Turning Point	NT2	0.3388	t=-0.338, p=0.737, p>.05
	NT3	0.319	
	Overt T2	0.2064	t=3.462, p=0.0014, p<.05
	Overt T3	0.4485	
	All T2	0.2267	t=3.466, p=0, p<.05
	All T3	0.3288	
Offset Pitch	NT2	138.97	t=1.458, p=0.153, p>.05
	NT3	144.85	
	Overt T2	154.4	t=-2.204, p=0.034, p<.05
	Overt T3	138.35	
	All T2	152.03	t=-3.005, p=0.0036, p<.05
	All T3	144.36	

Participant CBR: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	11.88	t=0.527, p=0.601, p>.05
	NT3	16.62	
	Overt T2	8.869	t=3.224, p=0.002, p<.05
	Overt T3	25.71	
	All T2	10.55	t=2.316, p=0.023, p<.05
	All T3	16.17	
Duration to the Turning Point	NT2	0.15	t=4.289, p=0, p<.05
	NT3	0.29	
	Overt T2	0.11	t=5.685, p=0, p<.05
	Overt T3	0.39	
	All T2	0.176	t=8.496, p=0, p<.05
	All T3	0.336	
Offset Pitch	NT2	162.95	t=-3.06, p=0.004, p<.05
	NT3	139.39	
	Overt T2	154.95	t=-3.058, p=0.004, p<.05
	Overt T3	134.27	
	All T2	155.53	t=-4.983, p=0, p<.05
	All T3	137.77	

Participant FZT: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	40.05	t=2.735, p=0.008, p<.05
	NT3	75.55	
	Overt T2	35.73	t=3.523, p=0.0015, p<.05
	Overt T3	97.8	
	All T2	41.56	t=3.901, p=0, p<.05
	All T3	71.36	
Duration to the Turning Point	NT2	0.298	t=2.556, p=0.028, p<.05
	NT3	0.418	
	Overt T2	0.27	t=3.233, p=0.0032, p<.05
	Overt T3	0.45	
	All T2	0.28	t=3.955, p=0, p<.05
	All T3	0.41	
Offset Pitch	NT2	303.62	t=-0.932, p=0.356, p>.05
	NT3	303.4	
	Overt T2	222.41	t=-3.967, p=0, p<.05
	Overt T3	284.45	
	All T2	318.31	t=-4.001, p=0, p<.05
	All T3	294.92	

Participant LKL: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	24.78	t=0.323, p=0.748, p>.05
	NT3	13.7	
	Overt T2	20.46	t=3.991, p=0, p<.05
	Overt T3	53.36	
	All T2	17.46	t=2.18, p=0.0319, p<.05
	All T3	41.07	
Duration to the Turning Point	NT2	0.37	t=-0.283, p=0.779, p>.05
	NT3	0.38	
	Overt T2	0.24	t=0.604, p=0.55, p>.05
	Overt T3	0.31	
	All T2	0.25	t=2.864, p=0.005, p<.05
	All T3	0.34	
Offset Pitch	NT2	210.09	t=0.208, p=0.837, p>.05
	NT3	211.47	
	Overt T2	222.93	t=-2.821, p=0.007, p<.05
	Overt T3	192.73	
	All T2	216.81	t=-2.006, p=0.0484, p<.05
	All T3	211.05	

Participant MXS: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	66	t=-1.344, p=0.2, p>.05
	NT3	35.17	
	Overt T2	39.65	t=8.666, p=0, p<.05
	Overt T3	187.64	
	All T2	35.35	t=5.54, p=0, p<.05
	All T3	135.34	
Duration to the Turning Point	NT2	0.59	t=-2.55, p=0.023, p<.05
	NT3	0.43	
	Overt T2	0.27	t=10.129, p=0, p<.05
	Overt T3	0.57	
	All T2	0.28	t=6.77, p=0, p<.05
	All T3	0.51	
Offset Pitch	NT2	191.49	t=1.08, p=0.298, p>.05
	NT3	232.44	
	Overt T2	246.05	t=-7.673, p=0, p<.05
	Overt T3	195.23	
	All T2	237.92	t=-6.6, p=0, p<.05
	All T3	201.03	

Participant NJ: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and

T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	83.06	t=1.141, p=0.264, p>.05
	NT3	154.52	
	Overt T2	43.53	t=9.001, p=0, p<.05
	Overt T3	195.08	
	All T2	43.42	t=6.183, p=0, p<.05
	All T3	160.96	
Duration to the Turning Point	NT2	0.52	t=-0.602, p=0.552, p>.05
	NT3	0.54	
	Overt T2	0.34	t=4.504, p=0, p<.05
	Overt T3	0.56	
	All T2	0.36	t=3.199, p=0.002, p<.05
	All T3	0.49	
Offset Pitch	NT2	216.81	t=-0.347, p=0.731, p>.05
	NT3	204.74	
	Overt T2	221.08	t= -1.959, p=0.056, p>.05
	Overt T3	211.22	
	All T2	218.6	t=-1.893, p=0.062, p>.05
	All T3	211.86	

Participant WYZ: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	16.94	t=1.2, p=0.243, p>.05
	NT3	25.11	
	Overt T2	14.6	t=5.542, p=0, p<.05
	Overt T3	53.47	
	All T2	14.81	t=4.558, p=0, p<.05
	All T3	38.89	
Duration to the Turning Point	NT2	0.34	t=-0.252, p=0.8033, p>.05
	NT3	0.36	
	Overt T2	0.41	t=0.592, p=0.556, p>.05
	Overt T3	0.40	
	All T2	0.38	t=-0.142, p=0.88, p>.05
	All T3	0.37	
Offset Pitch	NT2	144.28	t=0.297, p=0.769, p>.05
	NT3	145.79	
	Overt T2	148.5	t=-3.208, p=0.0022, p<.05
	Overt T3	138.49	
	All T2	146.34	t=-1.584, p=0.117, p>.05
	All T3	142.76	

Appendix B: Reports of Individual Participant's acoustic measurements in Elicitation 2 Analysis 1

Participant KL: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	35.796	t=2.218, p=0.22, p>.05
	NT3	50.95869408	
	Overt T2	46.80638362	t=1.802, p=0.087, p>.05
	Overt T3	97.081685	
	All T2	40.87747726	t=1.32, p=0.191, p>.05
	All T3	54.50661645	
Duration to the Turning Point	NT2	0.429051246	t=0.345, p=0.731, p>.05
	NT3	0.44997258	
	Overt T2	0.449417524	t=-0.874, p=0.393, p>.05
	Overt T3	0.356683253	
	All T2	0.44426831	t=-0.052, p=0.959, p>.05
	All T3	0.442796478	
Offset Pitch	NT2	269.9011685	t=-0.662, p=0.511, p>.05
	NT3	265.2477447	
	Overt T2	249.6540201	t=0.505, p=0.619, p>.05
	Overt T3	261.8741433	
	All T2	260.5563308	t=0.651, p=0.517, p<.05
	All T3	264.9882369	

Participant HL: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	20.52	t=1.256, p=0.215, p>.05
	NT3	32.83	
	Overt T2	14.283	t=2.861, p=0.007, p<.05
	Overt T3	42.66	
	All T2	40.87747726	t=2.796, p=0.0065, p<.05
	All T3	54.50661645	
Duration to the Turning Point	NT2	0.361	t=-1.424, p=0.162, p>.05
	NT3	0.322	
	Overt T2	0.318	t=1.743, p=0.091, p>.05
	Overt T3	0.357	
	All T2	0.44426831	t=-0.255, p=0.799, p>.05
	All T3	0.442796478	
Offset Pitch	NT2	153.199	t=-0.636, p=0.528, p>.05
	NT3	151.736	
	Overt T2	154.251	t=0.657, p=0.516, p>.05
	Overt T3	159.668	

	All T2	260.5563308	t=-0.014, p=0.989, p>.05
	All T3	264.9882369	

Participant JK: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	13.742	t=0.373, p=0.711, p>.05
	NT3	17.499	
	Overt T2	17.232	t=2.959, p=0.006, p<.05
	Overt T3	102.123	
	All T2	16.3367	t=0.69, p=0.492, p>.05
	All T3	21.838	
Duration to the Turning Point	NT2	0.258	t=0.616, p=0.541, p>.05
	NT3	0.291	
	Overt T2	0.218	t=1.845, p=0.075, p>.05
	Overt T3	0.406	
	All T2	0.228	t=2.081, p=0.040, p<.05
	All T3	0.331	
Offset Pitch	NT2	160.6	t=-0.882, p=0.383, p>.05
	NT3	145.28	
	Overt T2	145.791	t=1.108, p=0.277, p>.05
	Overt T3	185.66	
	All T2	149.589	t=-0.202, p=0.841, p>.05
	All T3	147.352	

Participant LWQ: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	15.54	t=1.354, p=0.188, p>.05
	NT3	49.78	
	Overt T2	8.377579789	t=6.329, p=0, p<.05
	Overt T3	56.82828627	
	All T2	8.928	t=6.205, p=0, p<.05
	All T3	51.608	
Duration to the Turning Point	NT2	0.419	t=0.041, p=0.968, p>.05
	NT3	0.421	
	Overt T2	0.300272889	t=5.148, p=0, p<.05
	Overt T3	0.485515202	
	All T2	0.309	t=5.34, p=0, p<.05
	All T3	0.447	
Offset Pitch	NT2	102.56	t=0.965, p=0.344, p>.05
	NT3	116.94	
	Overt T2	125.052919	t=-0.743, p=0.461, p>.05
	Overt T3	119.3575323	
	All T2	123.322	t=-0.946, p=0.347, p>.05

	All T3	117.93	

Participant FMY: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	23.37	t=2.246, p=0.029, p<.05
	NT3	63.05	
	Overt T2	23.61	N/A
	Overt T3	N/A	
	All T2	23.55	t=4.353, p=0, p<.05
	All T3	63.046	
Duration to the Turning Point	NT2	0.403	t=0.235, p=0.815, p>.05
	NT3	0.413	
	Overt T2	0.298	N/A
	Overt T3	N/A	
	All T2	0.325	t=2.948, p=0.004, p<.05
	All T3	0.413	
Offset Pitch	NT2	253.351	t=-0.267, p=0.79, p>.05
	NT3	251.312	
	Overt T2	257.575	N/A
	Overt T3	N/A	
	All T2	256.492	t=-1.234, p=0.221, p>.05
	All T3	251.312	

Participant LJ: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	269.373	N/A
	NT3	122.646	
	Overt T2	7.848	t=8.089, p=0, p<.05
	Overt T3	21.824	
	All T2	8.145	t=8.16, p=0, p<.05
	All T3	29.669	
Duration to the Turning Point	NT2	0.501973434	N/A
	NT3	0.29104552	
	Overt T2	0.293	t=7.305, p=0, p<.05
	Overt T3	0.367	
	All T2	0.298	t=7.091, p=0, p<.05
	All T3	0.4682	
Offset Pitch	NT2	126.4547355	N/A
	NT3	129.5406934	
	Overt T2	146.3	t=-7.435, p=0, p<.05
	Overt T3	108.43	
	All T2	145.365	t=-7.383, p=0, p<.05
	All T3	108.314	

Participant FZT: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	41.85	t=2.669, p=0.011, p<.05
	NT3	83.24	
	Overt T2	43.754	t=3.052, p=0.0039, p<.05
	Overt T3	80.6	
	All T2	43.068	t=4.158, p=0, p<.05
	All T3	81.951	
Duration to the Turning Point	NT2	0.482	t=-0.159, p=0.875, p>.05
	NT3	0.474	
	Overt T2	0.401	t=12.986, p=0, p<.05
	Overt T3	0.521	
	All T2	0.43	t=1.917, p=0.058, p>.05
	All T3	0.497	
Offset Pitch	NT2	275.84	t=-2.745, p=0.0098, p<.05
	NT3	230.16	
	Overt T2	274.507	t=-2.792, p=0.007, p<.05
	Overt T3	240.025	
	All T2	274.985	t=-4.073, p=0, p<.05
	All T3	234.963	

Participant WYZ: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	43.22	t=1.405, p=0.168, p>.05
	NT3	67.73	
	Overt T2	31.93	t=2.987, p=0.004, p<.05
	Overt T3	87.17	
	All T2	36.5594	t=3.139, p=0.002, p<.05
	All T3	75.701	
Duration to the Turning Point	NT2	0.349	t=1.689, p=0.099, p>.05
	NT3	0.419	
	Overt T2	0.296	t=4.116, p=0, p<.05
	Overt T3	0.482	
	All T2	0.318	t=4.169, p=0, p<.05
	All T3	0.445	
Offset Pitch	NT2	166.819	t=--0.533, p=0.597, p>.05
	NT3	163.22	
	Overt T2	174.425	t=-0.66, p=0.513, p>.05
	Overt T3	167.485	
	All T2	171.304	t=-1.038, p=0.303, p>.05
	All T3	164.972	

Participant CBR: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	10.53	t=1.535, p=0.135, p>.05
	NT3	42.14	
	Overt T2	8.122	t= 7.302, p=0, p<.05
	Overt T3	86.296	
	All T2	12.95741	t= 5.524, p=0, p<.05
	All T3	53.69337	
Duration to the Turning Point	NT2	0.349	t=1.689, p=0.099, p>.05
	NT3	0.419	
	Overt T2	0.24025	t= 11.879, p=0, p<.05
	Overt T3	0.68594	
	All T2	0.2558236	t= 4.903, p=0, p<.05
	All T3	0.4122572	
Offset Pitch	NT2	166.819	t=-0.533, p=0.597, p>.05
	NT3	163.22	
	Overt T2	148.903	t=1.187, p=0.242, p>.05
	Overt T3	137.05	
	All T2	149.3054	t=2.39, p= 0.0193, p<.05
	All T3	134.1993	

Participant NJ: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	50.69	t= 1.761, p= 0.098611, p>.05
	NT3	82.15	
	Overt T2	29.35	t= 6.686, p=0, p<.05
	Overt T3	149.19	
	All T2	35.36669	t= 6.914, p=0, p<.05
	All T3	138.8729	
Duration to the Turning Point	NT2	0.45153	t= 0.70, p= 0.495, p>.05
	NT3	0.48384	
	Overt T2	0.44288	t= 2.157, p= 0.0351, p<.05
	Overt T3	0.50014	
	All T2	0.4453177	t= 2.317, p= 0.0232, p<.05
	All T3	0.4976254	
Offset Pitch	NT2	198.09	t= 1.524, p= 0.148, p>.05
	NT3	163.22241.45	
	Overt T2	200.03	t= 3.459, p= 0, p<.05
	Overt T3	282.64	
	All T2	199.4825	t= 3.932, p= 0, p<.05
	All T3	276.3005	

Participant MSX: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2

and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	N/A	N/A
	NT3	N/A	
	Overt T2	19.076	t= 9.045, p=0, p<.05
	Overt T3	113.029	
	All T2	19.0757	t= 8.662, p=0, p<.05
	All T3	109.892	
Duration to the Turning Point	NT2	NA	N/A
	NT3	NA	
	Overt T2	0.287	t= 7.878, p= 0, p<.05
	Overt T3	0.536	
	All T2	0.2870403	t= 7.837, p= 0, p<.05
	All T3	0.3792337	
Offset Pitch	NT2	NA	N/A
	NT3	NA	
	Overt T2	218.239	t= -2.888, p= 0.005, p<.05
	Overt T3	195.122	
	All T2	218.2385	t= 2.97, p= 0.00398, p<.05
	All T3	195.0902	

Participant LKL: Data for each variable of the pair of NT2 and NT3, overt T2 and T3, and All T2 and T3			
Variable	Pair	Mean (Hz)	Significance value p
F0 Difference	NT2	22.77	t=2.523, p=0.025, p<.05
	NT3	83.13	
	Overt T2	31.19	t= 5.788, p=0, p<.05
	Overt T3	139.68	
	All T2	30.326	t= 5.828, p=0, p<.05
	All T3	123.728	
Duration to the Turning Point	NT2	0.402	t=-1.316, p=0.211, p>.05
	NT3	0.279	
	Overt T2	0.319	t= 2.594, p=0, p<.05
	Overt T3	0.409	
	All T2	0.327	t= 1.436, p= 0.155, p>.05
	All T3	0.379	
Offset Pitch	NT2	234.761	t=-2.796, p=0.015, p<.05
	NT3	220.71	
	Overt T2	234.578	t= -6.228, p= 0, p<.05
	Overt T3	215,916	
	All T2	234.596	t=-6.786, p= 0, p<.05
	All T3	217.268	

Appendix C: Results of the linear models constructed for the acoustic measures of each of the six participants' longitudinal production.

Table C.1: The results of the linear model constructed for the three acoustic variables from Participant CBR's longitudinal productions.

		Estimate	Std. Error	t value	Pr(>|t|)	
F0 Difference	(Intercept)	11.882	12.315	0.965	0.338	
	SegmentTone 3	2.559	13.215	0.194	0.847	
	ElicitationTime2	-1.351	18.472	-0.073	0.942	
	SegmentTone 3:ElicitationTime2	29.361	19.757	1.486	0.142	
Duration		Estimate	Std. Error	t value	Pr(>|t|)	
	(Intercept)	0.15093	0.05665	2.664	0.00966	**
	SegmentTone 3	0.18154	0.06079	2.986	0.00394	**
	ElicitationTime2	0.24113	0.08498	2.838	0.00601	**
	SegmentTone 3:ElicitationTime2	-0.17296	0.09089	-1.903	0.06134	.
Offset		Estimate	Std. Error	t value	Pr(>|t|)	
	(Intercept)	162.960	10.326	15.781	<2e-16	***
	SegmentTone 3	-24.582	11.081	-2.218	0.0299	*
	ElicitationTime2	-10.131	15.489	-0.654	0.5153	
	SegmentTone 3:ElicitationTime2	4.971	16.567	0.300	0.7651	
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*'					

Table C.2: The results of the linear model constructed for the three acoustic variables from Participant FZT's longitudinal productions.

		Estimate	Std. Error	t value	Pr(>|t|)	
F0 Difference	(Intercept)	40.059	9.002	4.450	2.78e-05	***
	SegmentTone3	25.524	11.139	2.291	0.0246	*
	ElicitationTime2	1.786	13.395	0.133	0.8943	
	SegmentTone3:ElicitationTime2	15.866	17.069	0.930	0.355	
Duration		Estimate	Std. Error	t value	Pr(>|t|)	
	(Intercept)	0.29872	0.03473	8.601	5.85e-13	***
	SegmentTone3	0.09379	0.04298	2.182	0.032053	*
	ElicitationTime2	0.18418	0.05168	3.564	0.000624	***
	SegmentTone3:ElicitationTime2	-0.10208	0.06586	-1.550	0.125120	
Offset		Estimate	Std. Error	t value	Pr(>|t|)	
	(Intercept)	303.630	8.528	35.603	<2e-16	***
	SegmentTone3	-6.416	10.553	-0.608	0.5450	
	ElicitationTime2	-27.788	12.690	-2.190	0.0315	*
	SegmentTone3:ElicitationTime2	-39.269	16.171	-2.428	0.0174	*
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*'					

Table C.3: The results of the linear model constructed for the three acoustic variables from Participant LKL's productions.

		Estimate	Std. Error	t value	Pr(>|t|)
F0 Difference	(Intercept)	24.787	43.128	0.575	0.568
	SegmentTone3	15.620	44.278	0.353	0.726
	ElicitationTime2	-2.017	52.820	-0.038	0.970
	SegmentTone3:ElicitationTime2	44.743	56.822	0.787	0.435
Duration		Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	0.36577	0.09114	4.013	0.0002 ***
	SegmentTone3	-0.02374	0.09357	-0.254	0.8007
	ElicitationTime2	0.03719	0.11162	0.333	0.7404
	SegmentTone3:ElicitationTime2	-0.09954	0.12008	-0.829	0.4110
Offset		Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	210.098	8.445	24.880	<2e-16 ***
	SegmentTone3	1.945	8.670	0.224	0.8234
	ElicitationTime2	24.663	10.342	2.385	0.0209 *
	SegmentTone3:ElicitationTime2	-15.996	11.126	-1.438	0.1567
	Signif. codes: 0 '***' 0.01 '*'				

Table C.4: The results of the linear model constructed for the three acoustic variables from Participant MXS's productions.

		Estimate	Std. Error	t value	Pr(>|t|)
F0 Difference	(Intercept)	66.00	20.06	3.291	0.00495 **
	SegmentTone3	-24.05	21.44	-1.122	0.27964
	ElicitationTime2	9.92	21.44	0.463	0.65022
	SegmentTone3:ElicitationTime2	NA	NA	NA	NA
Duration		Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	0.59047	0.09027	6.541	9.34e-06 ***
	SegmentTone3	-0.21437	0.09650	-2.221	0.0421 *
	ElicitationTime2	0.12845	0.09650	1.331	0.2031
	SegmentTone3:ElicitationTime2	NA	NA	NA	NA
Offset		Estimate	Std. Error	t value	Pr(>|t|)
	(Intercept)	191.49	16.66	11.492	7.79e-09 ***
	SegmentTone3	19.92	17.81	1.118	0.281
	ElicitationTime2	-16.90	17.81	-0.949	0.358
	SegmentTone3:ElicitationTime2	NA	NA	NA	NA
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*'				

Table C.5: The results of the linear model constructed for the three acoustic variables from Participant NJ's productions.

		Estimate	Std. Error	t value	Pr(>|t|)	
F0 Difference	(Intercept)	52.826	40.267	1.312	0.197	
	SegmentTone3	58.150	42.614	1.365	0.180	
	ElicitationTime2	-2.138	45.427	-0.047	0.963	
	SegmentTone3:ElicitationTime2	-26.689	55.397	-0.482	0.633	
Duration		Estimate	Std. Error	t value	Pr(>|t|)	
	(Intercept)	0.51952	0.09313	5.578	1.72e-06	***
	SegmentTone3	-0.07010	0.09856	-0.711	0.481	
	ElicitationTime2	-0.06799	0.10506	-0.647	0.521	
	SegmentTone3:ElicitationTime2	0.10241	0.12812	0.799	0.429	
Offset		Estimate	Std. Error	t value	Pr(>|t|)	
	(Intercept)	216.814	22.920	9.459	7.31e-12	***
	SegmentTone3	-5.502	24.257	-0.227	0.822	
	ElicitationTime2	-18.725	25.858	-0.724	0.473	
	SegmentTone3:ElicitationTime2	48.859	31.533	1.549	0.129	
	--- 0.0174 *					
	Signif. codes: 0 '***'					

Table C.6: The results of the linear model constructed for the three acoustic variables from Participant WYZ's productions.

		Estimate	Std. Error	t value	Pr(>|t|)	
F0 Difference	(Intercept)	16.95	24.91	0.680	0.499	
	SegmentTone3	9.44	26.63	0.355	0.724	
	ElicitationTime2	26.27	27.14	0.968	0.337	
	SegmentTone3:ElicitationTime2	15.07	30.10	0.501	0.619	
Duration		Estimate	Std. Error	t value	Pr(>|t|)	
	(Intercept)	0.340801	0.063833	5.339	1.57e-06	***
	SegmentTone3	-0.011569	0.068241	-0.170	0.866	
	ElicitationTime2	0.008579	0.069561	0.123	0.902	
	SegmentTone3:ElicitationTime2	0.081577	0.077151	1.057	0.295	
Offset		Estimate	Std. Error	t value	Pr(>|t|)	
	(Intercept)	144.283	10.324	13.975	<2e-16	***
	SegmentTone3	2.148	11.037	0.195	0.8464	
	ElicitationTime2	22.535	11.251	2.003	0.0498	*
	SegmentTone3:ElicitationTime2	-5.742	12.478	-0.460	0.6471	
	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Appendix D: Word list for tone production

Targets

- bá
- bă
- bí
- bĭ
- bú
- bŭ
- pí
- pĭ
- pú
- pŭ
- má
- mă
- mí
- mĭ
- fá
- fă
- fú
- fŭ
- dá
- dă
- dí
- dĭ
- dú
- dŭ
- tí
- tĭ
- tú
- tŭ
- ní
- nĭ
- nú
- nŭ
- ná
- nă
- lí
- lĭ
- lú
- lŭ
- hú
- hŭ
- jí
- jĭ
- qí
- qĭ
- xí
- xĭ
- gé
- gĕ
- ké
- kĕ
- zhé
- zhĕ
- shé
- shĕ
- zhí
- zhĭ
- zhú
- zhŭ
- chí
- chĭ
- chú
- chŭ
- shú
- shŭ
- shí
- shĭ
- shá
- shă
- rú
- rŭ
- yí
- yĭ
- yá
- yă
- yű
- yŭ
- wú
- wŭ

Fillers

- wèi
- wēi
- sī
- kāng
- xīn
- sù
- sū
- sān
- sàn
- tuō
- tuò
- chù
- chūqǜ
- shānshàng
- shàngzi
- xìngrén
- xīng
- lài
- shēng
- yào
- yāo
- le
- shēn
- shèn
- chuāng
- chuàngzào
- liàn
- bēi
- bèi
- zhuōbiān
- jiā
- jià
- kù
- àikū
- kuāng
- kuàng
- guān
- guàn
- kuān
- màozi
- bāobāo
- bàojǖn
- hūnyīn
- tàitài
- diànhuà
- fāng
- fàng
- gānjìng
- jīngyà
- bālè
- xiāngxìn
- bīng
- bìng
- jūn
- ēn
- jiāng
- jiàng
- chāo
- jvàn
- jvān
- tīng
- dìng
- dīng
- huì
- huī
- sàihuì
- sāi
- fēn
- fèn
- dānshù
- dàn
- guì
- guī
- hēng
- hèn
- ruì
- ōu
- gàosù

CURRICULUM VITAE

Li-Ya Mar (Maria)馬麗雅

Department of Linguistics, University of Wisconsin-Milwaukee,
Milwaukee, WI, 53201. liyamar@uwm.edu

EDUCATION

2016 **Ph.D.**, Linguistics. University of Wisconsin Milwaukee (UW-Milwaukee). Dissertation title: Covert contrast: The acquisition of Mandarin Tone 2 and Tone 3 in L2 production and perception. Advisor: Fred Eckman

2012 **M.A.**, Linguistics. University of Wisconsin Milwaukee. Master's paper: Tonal Adaptation of English Loanwords in Mandarin: The Role of Perception and Factors of Characters. Advisor: Gregory Iverson.

2010 **B.A**., Foreign Languages and Literatures. National Chiao Tung University (國立交通大學) (NCTU), Taiwan.

2010 **Certificate**, Teaching Chinese as a Second Language. NCTU, Taiwan

UNIVERSITY PROFESSIONAL EXPERIENCE

Senior Instructor, Center for Language Studies, Beloit College, Beloit, WI Summer 2014 & Summer 2015
- Coordinate and teach 3rd year Chinese in the immersion-style summer language program
- Design and teach intensive advanced-mid Chinese curriculum
- Involve teamwork with the program director, coordinator and other instructors

Graduate Teaching Assistant (Graduate Instructor), Linguistics, UW-Milwaukee August 2014-May 2016
- Stand-alone courses
- Teach introductory linguistic courses in English
- Face-to-face and online sections
- Design curriculum, plan lecture materials and arrange class activities
- Employ technology in classroom instruction
- Involve teamwork with the program director, coordinator and other instructors

Graduate Teaching Assistant (Graduate Instructor), Chinese, UW-Milwaukee August 2010-December 2015
- Stand-alone courses
- Teach post-secondary Chinese at the all levels (1st-8th semester Chinese, i.e. Chinese 101, 102, 201, 202, 301, 302, 401, and 402)
- Design curriculum, lesson plans, lecture materials and class activities
- Design assessment tasks of students' functional proficiency according to the ACFTL OPI guidelines
- Frequently employ technology in classroom instruction
- Involve teamwork with the program director, coordinator and other instructors
- Prepare materials for Chinese Language Conversation Table

Graduate Research Assistant, Linguistics**,** UW-Milwaukee August 2012-May 2013
- Research assistant in the phonetics lab
- Research topics include second language acquisition and language attitude
- Edit and analyze sound data on *Praat* and *Audacity*
- Record speech production of second language learners
- Conduct speech experiments on *E-Prime*

Graduate Project Assistant, Language Resource Center, UW-Milwaukee December 2012-May 2013
- Develop the online curriculum of first year Chinese
- Record course videos for first year Chinese online curriculum
- Edit course videos for first year Chinese online curriculum
- Create and correct online quizzes and lesson tests for first year Chinese online curriculum

Coordinator, Taiwanese Film Series, UW-Milwaukee December 2010
- Introduce modern Chinese film culture to the university community
- Arrange film screenings

- Lead post-screening discussions

Teaching Assistant, Mandarin, NCTU January-June 2010 & Summer 2011
- Drill and organize in-class activities
- Tutor students of Chinese
- Involve great amounts of teamwork with other instructors
- Correct students' assignments and exams

CONFERENCE PRESENTATIONS

Mar, L. Covert Contrast: The Acquisition of Mandarin T2 and T3 in L2 Production. The ACTFL Annual Convention and World Languages Expo. San Antonio, Texas. Paper Presentation. November 2014.

Mar, L. Covert contrast: The acoustic characteristics of Mandarin Tone 2 and Tone 3 in L2 production. The 33rd Second Language Research Forum. Columbia, South Carolina. Paper Presentation. October 2014.

Mar, L. Covert contrast in L2 acquisition: The acoustic characteristics between Mandarin Tone 2 and Tone 3. The 2014 Meeting of the Graduate Workshop of the American Midwest and Prairies. Milwaukee. Paper Presentation October 2014.

Mar, L. Tonal Adaptation of the Stress in Initially-Stressed Disyllabic English Loanwords in Mandarin Chinese: Effects of Vowel Type and Onset Consonants. The 21st Annual Meeting of the International Association of Chinese Linguistics. Taipei, Taiwan. Paper Presentation. June 2013.

Park, H. and Mar, L. English influence as a possible source for the sound change in Korean. Mid-Continental Phonetics & Phonology Conference 18. Ann Arbor, Michigan. Poster Presentation. March 2013.

Mar, L. and Park, H. Tonal Adaptation of English Loanwords in Mandarin: The Role of Perception and Factors of Characters. The 164th Meeting of the Acoustical Society of America. Kansas City, Missouri. Poster Presentation. Oct 2012.

OTHER PROFESSIONAL EXPERIENCE

Overseas Volunteer English Teacher, University System of Taiwan Summer 2008 & Summer 2009
- Teach English at the high school level in Ningxia Autonomous Region, China
- Intensive program: 4 hours per day
- Organize lesson plans and teaching groups
- Involve a lot of teamwork with other volunteers and school administrators

Coordinator, English Lunchtime **(**chat room for improving English speaking skills), NCTU March 2007-June 2009
- Prepare discussion materials and lead discussion
- Involve teamwork with other program planners and school administrators
- Communicate with the university community members with diverse background
- Selected by program faculty

Coordinator, English Fun Time (chat room for improving English speaking skills), Industrial 2008-2009
Technology Research Institute of Taiwan (台灣工業技術研究院)
- Prepare discussion materials and lead discussion
- Involve a lot of teamwork with other program planners and school administrators
- Communicate with the engineers with diverse disciplines

English Teacher, Pu-Tai International English Summer Camp, Pu-Tai Elementary Summer 2007
and Junior High School, Puli, Taiwan
- Duties include: pre-class preparation, after-class review, facilitate communication between native teachers and students, take care of students in activities
- Involve teamwork with other English-speaking TAs and school administrators

Volunteer Online English Tutor, NCTU 2009-2010
- Tutor typhoon victims in Jia-Xian Junior High School, Kaohsiung, Taiwan

PROFESSIONAL DEVELOPMENT

OPI (Oral Proficiency Interview) Workshop: Crossing Major Borders, UW-Milwaukee September 2012
- One-day workshop on how to help students cross proficiency level (Trainer: Dr. Karl F. Otto, Jr)

OPI Familiarization Workshop, UW-Milwaukee March 2012
- One-day workshop on OPI standards

ASA School, Acoustical Society of America, Kansas City, Missouri October 2012
- Three-day workshop on a wide variety of topics related to the interdisciplinary acoustical theme *Living in the Acoustic Environment*

Startalk Program, University of Minnesota-Twin Cities June-July 2011
- Complete 60-hour professional development course entitled "Culture as the Core in the Second Language Classroom (exclusively for Mandarin teachers)

ACTFL Annual Convention and World Languages Expo, volunteer November 2011
- Chinese Language Teacher Association (CLTA)
- Assist academic scholar present significant findings of their studies
- Coordinate annual member meeting and conference dinner

ACADEMIC AWARDS

Chancellor's Graduate Student Award, UW-Milwaukee (USD$1,000), 2012-2013

Chancellor's Graduate Student Award, UW-Milwaukee (USD$1,000), 2014-2015

PROFESSIONAL ORGANIZATIONS

Chinese Language Teacher Association

Linguistics Society of America

International Association of Chinese Linguistics

Acoustical Society of America

The Wisconsin Association of Chinese Language Teacher

LANGUAGES

Mandarin Chinese (native)

English (near-native)

Tiawanese (intermediate)

Spanish (rudimentary)

REFERENCES

Avalable upon request