2008

# Development of Lexical Tone Production in Disyllabic Words by 2- to 6-year-old Mandarin-speaking Children

Puisan Wong
*The Graduate Center, City University of New York*

# Development of Lexical Tone Production in Disyllabic Words by 2- to 6-year-old Mandarin-speaking Children

by

PUISAN WONG

A dissertation submitted to the Graduate Faculty in Speech-Language-Hearing Sciences
in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York

2008

This manuscript has been read and accepted for the Graduate Faculty in Speech-Language-Hearing Sciences in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

| | |
|---|---|
| Date | Winifred Strange, Ph.D.<br>Chair of Examining Committee |
| Date | Martin Gitterman, Ph.D.<br>Executive Officer |
| | Laura Koenig, Ph.D. |
| | John Locke, Ph.D. |
| | Supervisory Committee |
| | Susan Nittrouer, Ph.D. |
| | Outside Reader |

THE CITY UNIVERSITY OF NEW YORK

# Abstract

DEVELOPMENT OF LEXICAL TONE PRODUCTION IN DISYLLABIC WORDS BY 2- TO 6-YEAR-OLD MANDARIN-SPEAKING CHILDREN

by

Puisan Wong

Advisor: Professor Winifred Strange

This study investigated children's development in the production of Mandarin lexical tones in familiar disyllabic words and tested the hypothesis that disyllabic tone contours with more complex fundamental frequency contours are more difficult for children to produce. Participants were forty-four 2- to 6-year-old monolingual Mandarin-speaking children and 12 mothers. Their disyllabic tone productions were elicited by picture naming and low-pass filtered to eliminate lexical information while retaining the fundamental frequency contours. Three Mandarin-speaking judges listened to the filtered stimuli, and categorized the children's and adult's disyllabic tones. Acoustic analysis was performed on selected accurate child and adult productions and on a sample of children's inaccurate productions.

Judges identified adults' productions as the intended tones with very high accuracy. As a group, children's productions were judged significantly less correctly than adults'. Judged correctness increased significantly with age, but even 5- to 6-year-old children's disyllabic tones were judged as less accurate overall than adults'. Large inter-subject variability was observed in 2- to 4-year-old children's performance. Some

disyllabic tones, particularly non-compatible tone combinations (i.e., tones with large transitions at the boundary between the syllables), remained difficult even for older children. When children made errors, they usually produced one of the tones correctly; error patterns suggested that they modified the first tone to be more compatible with the second tone (i.e., showed anticipatory coarticulation patterns), unlike the adult patterns which show more carry-over coarticulatory effects. When the four lexical tones were analyzed separately, significant context effects were found. Children produced the high level tone (T1) more accurately in the second than the first syllable. The rising tone (T2) was more accurately produced in compatible than non-compatible contexts. The low, dipping tone (T3) and falling tone (T4) were produced least accurately in the first syllable when the tone combination was non-compatible.

In conclusion, acquisition of disyllabic Mandarin tone contour appears to be a gradual process that spans more than six years to achieve mastery. Children have more difficulty producing complex tone contours that demand rapid f0 changes, suggesting the influence of immature physiological control of laryngeal gestures on the production of lexical tone contours in continuous speech.

# Dedication

獻給我的父母

# Acknowledgments

I would like to express my deepest thanks and appreciation to many people who have made my graduate studies at the Graduate Center of the City University of New York such a rich, joyful and memorable experience in my life. I have been very fortunate to have Dr. Winifred Strange as my advisor and mentor. Over the years I have been inspired by her passion, knowledge and seriousness in science. I benefited immensely by listening to her enthusiastic discussions on different ongoing research projects in the weekly project meetings and the wide range of interesting presentations in the lab meetings. Her devotion to the students was admirable. She treated her students with respect and was very generous with her time and expertise. Each of my individual meetings with her was stimulating, insightful and productive. With her sharpness, keen insight and passion, she always amazed me by how much we could accomplish in each meeting. She set very high standards for the students and for research but at the same time was very encouraging and understanding. Without her full support, sacrifice and encouragements, finishing my dissertation would have been impossible. I thank her from the bottom of my heart for making my doctoral training truly rewarding, exciting and enjoyable.

I had a remarkable dissertation committee. I am very grateful to Dr. Laura Koenig for her insightful input. Her thoughtfulness and kindness during this process supported me profoundly. I felt very privileged to have Dr. John Locke serving on my committee. He generously contributed his time and expertise to my study and was very supportive and encouraging. It was my great pleasure to have Dr. Susan Nittrouer as my outside

reader. I was greatly stimulated by her helpful comments and enormously encouraged by her kind words.

I am extremely grateful to Dr. Yi Xu for his generosity, his expertise, and his continual assistance and support. I thank Gary Chant, Bruno Tagliaferri, and Marcin Wroblewski for their valuable and unfailing technical support. I thank William Hui for the magnificent pictures and NIH for the financial support.

I am in great debt to the participants, their families, and the individuals who helped recruiting participants for the study. Many of them touched me with their great enthusiasm and immense interest in supporting developmental research. I particularly would like to thank Dr. Josephine Jung, Dr. Elsie Lee, Shari Cai, Yvette Chen, Ivy Xu, Cheryl Cai, and Iris for their generous help in recruiting children.

I extend my gratitude to Dr. Martin Gitterman for his kindness and warm support in different aspects of my doctoral studies. I thank Dr. James Jenkins for his guidance and support and being an excellent role model; his wisdom, generosity, and kindness will be among the fondest memories of my graduate studies. I thank Dr. Loraine Obler for being a wonderful mentor and strong source of support in various aspects of my life. I thank Dr. Richard Schwartz heartily for encouraging me to pursue my doctoral studies and instilling my earliest research interests.

I want to thank all the faculty and staff in the department for cultivating such an intellectual, stimulating, nurturing, caring, warm and homey atmosphere in the department, which I will miss tremendously. I would like to thank Loretta Walker for her administrative assistance. I have been blessed to have met so many bright, interesting, and receptive students in the program. They have been my source of inspiration, joy,

laughter and support. The friendships I have developed will continue to be precious and dear to me.

Most of all I thank my family for their love, understanding and support. Without them all these endeavors would have been impossible. I owe a huge debt of gratitude to my parents who did not have the privilege to receive higher education but have been strong supporters and believers in education. To them I dedicate this dissertation. I thank my brother George for his understanding and emotional support. I thank my husband Xiliang for his unconditional support in everything I do, in going through the ups and downs in life with me, and for his love, patience, understanding, and encouragement. This journey would not have been such a pleasure without him on my side.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1: Introduction

## Statement of Problem

Lexical tone is an essential component of the phonological structure in over 60% of the world's languages (Yip, 2002). However, little is known about children's tone development. Previous research has sometimes suggested that tones are acquired very early before segmental speech sounds are mastered. However, a recent study found that children had not mastered the four Mandarin tones in monosyllabic words by the age of three years (Wong, Schwartz, & Jenkins, 2005). Due to limited data and conflicting results, tone is usually neglected in phonological theories or theories of phonological development despite the importance of tone in most languages. This study tracks the developmental process of the production of Mandarin lexical tones in familiar disyllabic words by 2- to 6-year-old Mandarin-speaking children.

## Introduction to Tones in Mandarin

Mandarin, the most widely spoken language in the world, uses tone to make lexical contrasts. Mandarin has four full tones: Tone 1 (T1), Tone 2 (T2), Tone 3 (T3), and Tone 4 (T4) (Chao, 1968; Li & Thompson, 1989). When produced in isolation, the four tones have a high level (H), mid rising (R), low dipping (falling rising) (L), and high falling (F) fundamental frequency (f0) contour, respectively. Syllables with the same phonetic segments but a different tone have distinct meanings. For example, when produced in the four tones, the syllable 'ma' means 'mother', 'hemp', 'horse' and 'scold', respectively. Mandarin has a fifth tone, the neutral tone, which is found in weakly

stressed syllables such as suffixes (e.g., aspect markers), particles (e.g., question particles), and the second syllable of reduplicated verbs or reduplicated kinship terms. Some studies suggest that the neutral tone has no specific f0 contour (Yip, 2002); its f0 value varies according to the f0 values of the preceding tone (Chao, 1968; Shen, 1992). Other studies suggest that the neutral tone has a mid and static f0 target (Chen & Xu, 2006).

The primary and sufficient cue for Mandarin tone recognition is the fundamental frequency contour (Fu & Zeng, 2000; Luo & Fu, 2004; Massaro, Cohen, & Tseng, 1985; Whalen & Xu, 1992), although other acoustic features such as vowel duration (Fu & Zeng, 2000; Ho, 1976; Shen, 1990; Xu, 1997), syllable amplitude (Gårding, Kratochvil, Svantesson, & Zhang, 1986; Howie, 1976; Whalen & Xu, 1992) and vocal quality (e.g., creaky voice) (Gårding et al., 1986) are found to covary with f0 change in Mandarin tones. In the presence of f0 cues, other cues are negligible (Fu & Zeng, 2000; Gårding et al., 1986; Whalen & Xu, 1992)

Figure 1 shows the mean f0 contours of 48 tokens of each of the four Mandarin tones in the syllable /ma/ produced in isolation by eight adult males (Xu, 1997). When produced in connected speech, T3 undergoes phonological changes. It becomes a low level tone in non-final positions and a rising tone when preceding another T3. The latter (i.e., T3 + T3 → Rising + T3) is known as the T3 sandhi rule. The Rising + T3 as a result of applying the sandhi rule to T3T3 combinations is often perceived by native speakers as the same as T2T3 (Wang & Li, 1967; Pang 2000, cited in (Xu, 2005)).

## Target Approximation Model of Mandarin Tones

Xu (1997; 1999; 2001) proposed a Target Approximation Model (TAM) to characterize adults' perception and production of Mandarin tones. In this model, the syllable is the production unit of tone. That is, the articulatory gestures for the tone start at the beginning and terminate at the end of the syllable. Each tone has an underlying pitch target (i.e., an ideal pitch pattern); these are high (H), rise (R), low (L), fall (F) for the four tones, respectively. These underlying pitch targets are the articulatory goals for the production of the four tones. However, the same underlying articulatory goal may not have exactly the same surface f0 because during production, the implementation of the underlying pitch target in different contexts is often affected by articulatory constraints such as maximum speed of pitch rise and pitch fall, duration of the syllable for implementing the tone, inertia of speech movements (articulatory movements cannot stop or change direction instantaneously), and the state (e.g., velocity and displacement) of the articulators at the beginning of the syllable. Given the articulatory constraints, the degree of approximation of the pitch target varies in different tonal contexts and the pitch target is best approximated at the end of the syllable.

In connected speech, tones are produced in sequence and the f0 contours are substantially affected by syllable position and the preceding tone (Xu & Wang, 2001; Xu, 2001). When producing a disyllabic word, the first syllable (S1) is primarily the host for the f0 contour of the first tone and there is little effect of the second tone on the first tone (small anticipatory effect). However, the f0 contour of the second syllable (S2) is substantially affected by the tone of S1 (large carryover effect). Essentially, the initial portion of S2 is the transition from the f0 offset of S1 to the onset f0 of the target tone in

S2. The f0 transition can take more than 2/3 of the duration of S2. Thus, the same tone produced in S2 tends to have a more complex surface f0 contour than when it is produced in S1. In addition, the carryover effect of S1 on S2 is much greater when the f0 offset of S1 is very different from the f0 onset of S2 (i.e., non-compatible conditions) than when the f0 offset of S1 is similar to the f0 onset of S2 (i.e., compatible conditions). For instance, T4 (F) is in a compatible (C) context when preceded by T1 (H) but in a non-compatible (NC) context when preceded by T3 (L). Table 1 lists all the C and NC tone combinations in disyllabic words. Because T3T3 becomes T2T3, it is not listed in the table. Figure 2 is a reprint of figure 2 in (Xu, 2001). It shows the f0 contours of the 16 combinations of the four Mandarin tones. Each contour represents 48 productions of the tone combinations in the syllable /mama/ by eight native male speakers. When comparing the f0 contours of the same tone in S2 in C versus NC contexts, the f0 at the beginning of S2 in NC contexts is further away from the pitch target and it usually involves a change of the direction of the f0 contour to achieve the target tones. Thus, the f0 contour of the same tone is more complex when produced in S2 in NC contexts than in S2 in C contexts. Taken together, there seems to be a hierarchy of complexity of f0 contours as a function of contexts in disyllabic word production. The tonal contexts in the order of increasing f0 complexity are: S1, S2 in compatible contexts, and S2 in non-compatible contexts.

## Tone Production and Physiological Constraints

More complex f0 contours (e.g., f0 contours of S2 in NC contexts) are more difficult to produce than less complex f0 contours (e.g., f0 contours of S2 in C contexts) due to physiological constraints in tone production. Tone is implemented within the time frame of a syllable (Xu & Wang, 2001). Thus, the production of more complex f0

contours requires a greater degree of f0 change (e.g., higher velocity and more changes of f0 direction) within the syllable frame than less complex f0 contours. Yet there are physiological limits on the maximum speed of pitch change (i.e., acceleration and deceleration of f0) by the articulators, and the maximum speed of pitch change is often approached by the adult speaker during speech production (Xu & Sun, 2002). Therefore, the production of more complex f0 contours imposes greater demands on the laryngeal system and is more susceptible to target undershoot (Xu, 2005).

Physiologically, the production of tone involves graded/gradient control, precise coordination and constant adjustment of the laryngeal musculature to modify the length and tension of the vocal folds (Seikel, King, & Drumright, 1997; Titze, 1994; Zemlin, 1988). For example, a rising tone is produced by stretching the vocal folds so that they become thinner and stiffer. This results in an increase in the rate of vibration of the vocal folds, which, in turn, gives rise to a percept of higher pitch. On the other hand, if the vocal folds are shortened, thickened, and/or slackened, the frequency of vibration of the vocal folds decreases, resulting in a percept of lower pitch (Titze, 1994). Thus, to produce accurate Mandarin tones, children have to control and coordinate the laryngeal muscles to increase and decrease f0 precisely and efficiently.

Young children show significant differences from adults in their speech motor control (Smith, 2006) and anatomy (Kent & Vorperian, 1995), and do not possess the adult-like laryngeal physiology for tone production. First (Kent & Vorperian, 1995), the larynx in infants and children is still undergoing anatomical changes (Bosma, 1975; Crelin, 1987; Hirano, Kurita, & Nakashima, 1981; Kahane, 1978; Kahane, 1982; Kent & Vorperian, 1995). The growth of the larynx is particularly rapid between birth and 18

months of age (Tucker & Tucker, 1979 cited in (Kent & Vorperian, 1995) and continues to undergo considerable development until 5 or 6 years of age (Crelin, 1987; Hirano et al., 1981).The laryngeal cartilages, muscles, mucous membranes, and submucosal tissues become firmer and less pliable as the child gets older (Kent & Vorperian, 1995; Titze, 1989). Infants' vocal folds are about 1/6 to 1/3 of the length of the adults' (Hirano et al., 1981; Kent & Vorperian, 1995) and lack the distinctive layered structure found in adults' vocal folds (Hirano et al., 1981; Hirano et al., 1983). The length of the vocal folds continues to increase until around 20 years of age (Hirano et al., 1981; Kent & Vorperian, 1995) and the composition of the laryngeal musculatures does not reach the adult form until 16 years of age (Kent & Vorperian, 1995).

Second, children are thought to produce slower speech movements than adults in general (Goffman & Smith, 1999; Smith, 2006; Smith, 1978). Three-year-old children tend to produce speech sounds with longer durations than five-year-old children (Walker, Archibald, Cherniak, & Fish, 1992). The articulation rates of three- to seven-year-old children are slower than those of older children's or adults' (Smith, 1991). Adult-like speech rates are not reached until 14 – 16 years of age (Smith, 2006).

Third, children are less mature in speech motor coordination, which involves temporal and spatial control of the articulatory musculature. Temporal control entails the activation and deactivation of the articulatory muscles at the right time. Spatial control involves the activation of the appropriate muscles and the appropriate subgroups of motor units within those muscles to produce finely graded muscle activities (Smith & Zelaznik, 2004). Young children's articulatory gestures are typically unstable and uncoordinated (Goodell & Studdert-Kennedy, 1993; Kleinow & Smith, 2006). Their speech productions

tend to be more varied than older children's and adults' (Goffman, Gerken, & Lucchesi, 2007; Kleinow & Smith, 2006; Walsh & Smith, 2002; Wohlert & Smith, 2002). For example, four- and six-year-old children showed longer and more variable productions than 12-year-olds and adults (Kent & Forner, 1980; Kent & Forner, 1980). Their oral motor coordination patterns are not adult-like even after 14 years of age (Smith, 2006). Thus, the development of speech motor control and coordination is a gradual process and takes an extended period of time to achieve the adult form.

Children's acquisition of speech is dependent on the maturation of the articulators and articulatory motor control (Kent, 1976; Kent, 1984; Kent, 1992; Locke, 1986). When children learn to produce Mandarin tones, not only do they need to acquire the accurate phonologic representation of the tone, but they also have to master sophisticated skills in coordinating the laryngeal muscles to regulate the tension of the vocal folds to increase and decrease f0 precisely and efficiently, in coordination with supralaryngeal articulatory gestures. Given that young children's larynges are not fully developed, their articulatory gestures are slower, and their control and coordination of the articulatory musculature are less mature, the development of tone production is expected to be a gradual process. It was hypothesized here that accuracy rates in tone production will increase with age and more complex tone combinations would be produced less accurately by children than less complex tone combinations.

## Previous Studies on Children's Acquisition of Mandarin Tones

Several studies have reported on children's development of Mandarin tone productions. Some were case studies that collected longitudinal data from one to four children (Chao, 1973/1951; Clumeck, 1977; Hua, 2002). Chao (1973/1951) reported the

phonological development of his 28-month-old granddaughter acquiring Mandarin in the United States. Based on her spontaneous productions, the child was reported to have mastered the tones at the onset of the study and displayed only some errors in tone sandhi rules. Thus, very little information was provided on the tone development of the child.

Clumeck (1977) reported longitudinal data for three children learning Mandarin as a first language in the U.S. One boy was followed from 14 to 32 months of age. His family spoke Shanghainese, a Chinese tonal dialect, and Mandarin. The child did not start producing single words until 18 months of age. He produced all words with a rising pitch at the age of 1;10 and started to produce H and F tones at 1;11. The tone system was not fully mastered at 32 months (Clumeck, 1977; Clumeck, 1977). Another boy and a girl were followed from 2;3 to 3;5 and 1;10 to 2;10, respectively. Their imitated, elicited and spontaneous productions of isolated words and words in utterance-final position were analyzed. The accuracy rates for the four tones (i.e., H, R, L, F) were 97%, 83.3%, 87.4%, 94.3% for the boy and 97.2%, 61.3%, 73.9% and 95.8% for the girl. The girl continued to demonstrate difficulty with L and R tones at 2;10 (Clumeck, 1980). For both children H and F tones were acquired before R and L tones.

Hua (2002) studied tone production of four children from 10 to 24 months in Beijing. Tone productions were collected in mother-child interactions during play. The researchers reported that all the children produced the tones accurately in their spontaneous speech before the age of two years. Moreover, H and F tones were reported to be produced relatively early (i.e., the child imitated or spontaneously produced the tone at least once), while the L tone was the last to emerge. H and F tones “stabilized” (i.e., were produced with an accuracy rate of 66.7% or more in the speech samples, as defined

by the authors) before R and L tones, and the H tone was frequently used as a substitute for the other tones.

Three larger-scale studies on children's production of Mandarin tones have been published (Hua & Dodd, 2000; Li & Thompson, 1977; Wong et al., 2005). Li and Thompson (1977) collected cross-sectional data on 17 children ages 1;6-3;0 in Taiwan. Longitudinal data were also collected for 10 of the children and intermittently with the other 7 children for 7 months. Tones were elicited by picture naming. Children's tone development was described as a 4-stage process. At Stage I, when children had a limited vocabulary and produced mostly single word utterances, H and F tones were predominantly produced. At Stage II, when the children had a larger vocabulary but were still producing mostly one-word utterances, they produced the four contrastive tones with occasional confusion of the R and L tones. At stage III, when the children produced predominantly 2 or 3 word utterances, tone sandhi rules emerged, but R and L tones continued to be produced less accurately. At Stage IV, when the children started to produce longer utterances, they produced all four tones. No data were provided on the stimuli, the age of mastery for the tones, or the age, the number, or the accuracy rates of the children in each stage of development.

Hua and Dodd (2000) examined the phonological development of 129 children ages between 1;6-4;6 in Beijing. Forty-four words of one to three syllables were elicited in a picture naming task and the children were also asked to describe four 5-scene pictures. Children's productions were transcribed by a phonetician. The researchers reported that although children made some errors in sandhi rules and the production of neutral tones, out of the whole database of 129 children, only 2 tone errors with the four

full tones were found, suggesting that children as young as 1;6 had mastered the production of the four tones in various contexts (Hua & Dodd, 2000; Hua, 2002)

All of the preceding studies of children's Mandarin tone production determined children's accuracy in tone productions based on natural/unprocessed speech with the support of lexical, contextual and/or linguistic information. Thus, judges' lexical and linguistic knowledge may have biased their perception of tone contours. The most recent study examined 13 three-year-old monolingual Mandarin-speaking children's production of tones in monosyllabic words using a picture naming task in the U.S. (Wong et al., 2005). Unlike previous studies, the productions of the children and four of their mothers were recorded and low-pass filtered to eliminate the segmental information, while preserving the f0 contours. Twelve native Mandarin-speaking judges were asked to categorize the tone patterns in the children's and adults' productions based on the filtered speech. The adults' tone productions were categorized with 96%, 96%, 83% and 98% accuracy, while children's production were categorized with 78%, 70%, 44%, and 76% accuracy for the 4 tones (H, R, L, F), respectively. The judges made significantly more errors in identifying children's than adults' tone productions. The L tones were more difficult to identify than any of the other three tones in both adults and children's productions. All adults' and most of the children's L tone errors involved the perceived substitution of the R tone for the L tone. Though similar error patterns (e.g., R and L tone confusions) were found in adults' and children's productions, children's error patterns were more diverse. Six of the 12 error patterns that occurred in the children's productions never occurred in adults' productions. These results indicate that children learning

Mandarin as their first language in the U.S. had not mastered the production of the four tones in monosyllabic words by the age of three years.

In summary, conflicting results have been found for the age and order of acquisition of Mandarin tones. Some studies suggested that tones were acquired very quickly and early; children produced the tones correctly in various contexts before the age of two years (e.g., (Hua & Dodd, 2000; Hua, 2002), whereas others found that children had not mastered the production of tones in monosyllabic words by the age of three years (Wong et al., 2005). Most studies found that H and F tones were acquired before R and L tones (Hua, 2002; Li & Thompson, 1977) and that most tone errors involved substitution between R and L tones (Clumeck, 1980; Li & Thompson, 1977; Wong et al., 2005). However, others reported that the R tone was acquired first (Clumeck, 1977), and that most of the errors involved using the H tone to substitute for other tones (Hua, 2002). Still others found that the L tone was the hardest (Wong et al., 2005).

The discrepancies in the findings about the age of tone mastery could be due to methodological differences. Most studies determined children's tone accuracy by only one judge (Chao, 1973/1951; Clumeck, 1977; Hua & Dodd, 2000; Hua, 2002) with the support of contextual, semantic, syntactic and segmental cues. In such cases it is impossible for the judge to eliminate the effect of his/her language skills and his/her tone expectations on the judgments, causing transcriber biases (S. Nittrouer, 1995; Oller & Eilers, 1975). The use of different criteria may also have contributed to the divergent findings. Some studies did not mention the criteria or the accuracy rates used (e.g., Chao, 1973/1951; Li & Thompson, 1977). Hua (2002) and Hua and Dodd (2000) set some criteria for emergence and stabilization: a tone was defined as "emerged" if 90% of the

children in the age group imitated or produced the tone at least one time and "stabilized" when 90% of the children in the age group produced the tone with 66.7% or higher accuracy rate (Hua & Dodd, 2000). Only one study directly compared children's productions to the adult forms (Wong et al., 2005). Most studies involved very few children, so the findings may not be representative. Though children's connected speech was collected in most studies, no studies have systematically examined children's production of coarticulated tones. As a result, it remains unclear when children produce adult-like Mandarin tones and how children's tone production changes with age.

There are two parts in the present study. The first part is a study of native Mandarin speakers' perception of tones in disyllables (hereafter DTs) produced by 2- to 6½-year-old children and the second part is an acoustic study on the children's disyllabic tone (DT) productions.

# Chapter 2: Perceptual Judgments of Children's Tone production

## Research Questions

In this study, we adopted the same methods used in Wong, et al., (Wong et al., 2005) to examine 2- to 6-year-old children's production of full Mandarin lexical tones in familiar disyllabic words. Neutral tones were not included. The first goal was to examine the accuracy rates of DTs in 2- to 6-year-old children. Specifically, we examined whether children's accuracy rates in DT productions changed as a function of age and how children approached the adult forms over time. Questions addressed included: (1) Were 2- to 6-year-old children's accuracy rates in productions of disyllabic tones (DT) adult-like? (2) Was there any developmental trend in the accuracy rates of children's DT productions? (3) What were the error patterns in children's tone productions?

The second goal of the perceptual judgment study was to test the hypothesis that children's tone accuracy in disyllables would reflect f0 complexity. As indicated above, the complexity of the f0 contours varies systematically in different contexts. Given that children tend to have immature speech anatomy and physiology, we hypothesized that children's accuracy rates in tone production would decrease as the f0 contours became more complex. Specifically, we predicted that (1) children's accuracy rates of DTs would be higher in compatible (C) than in non compatible (NC) tone combinations; and (2) children's tone accuracy rates would be higher in the first syllable (S1) than in the second syllable (S2). As part of this analysis, the accuracy of the four Mandarin tones (T1, T2, T3, T4) as a function of syllable position and compatibility of the tones preceding/following it was explored.

# Method

## Children's and Adults' Tone Productions

### Participants

Children. Seventy-five Mandarin-speaking children (44M, 31F, age range = 2;1-6;7) were recruited in the Tri-State Area of New York. Parents of the children filled out a questionnaire and provided information on the cognitive, social, physical, emotional, educational, speech and language backgrounds of the child. The children were given a Chinese speech and language test—Language Disorder Scale of Preschoolers (LDSP, 學前兒童語言障礙評量表) (Lin & Lin, 1994), an English language test—Preschool Language Scale-4 (PLS-4) (Zimmerman, Steiner, & Pond, 2002), and a hearing screening. A language sample was collected via play, story telling, story retelling and/or conversation. In order to determine that the child participants were normally developing and were native Mandarin learners with limited exposure to other languages or dialects, children had to meet the following criteria to be included in the study: (1) the child must have unremarkable cognitive, social, physical, emotional, educational and speech and language history according to parental report; (2) family members and caregivers spoke only Mandarin to the child; (3) the child scored higher than the 20$^{th}$ percentile rank in the total language score in LDSP and lower than the 20$^{th}$ percentile rank in the total language score in PLS-4; (5) no language limitations or atypicalities were observed in the language sample; (6) the child passed the hearing screening at 1K, 2K and 4K Hz at 20 dB HL under headphones using conditioned play audiometry (American Speech-Language-

Hearing Association, 1997); and (7) the child did not have a history of chronic otitis media according to parental report.

Thirty-one children failed to meet all the inclusion criteria and were excluded from the study. Five children had family members who spoke English to them. Three children were exposed to Japanese or Spanish. Eight children were exposed to another Chinese dialect in the family. Nine children had their total English score higher than the 20$^{th}$ percentile rank. Two children scored below the 20$^{th}$ percentile rank for both English and Chinese total scores. One child was suspected as having language delays. One child failed the hearing screening. Two children did not cooperate and failed to finish the tasks.

Thus, forty-four children (M = 17, F = 27, age range = 2;1-6;7) were included in the study. There were 12 two-year-old children (i.e., between 2;0-2;11), 13 three-year-old children (i.e., between 3;0-3;11), 11 four-year-old children (i.e., between 4;0-4;11) and eight children who were 5 years or older (i.e., between 5;0-6;7) (See Appendix A).

Appendix A shows the children's language scores and information on their deomographic backgrounds. The language tests were attempted but no language scores were provided for the youngest two children (UC66[1] and UC67, aged 2;1 and 2;2, respectively). The English test was discontinued because both children did not respond to the test items. Parents of the children reported no exposure to English. One of the two children (UC66) was in the U.S. for only two weeks. Neither child had the attention span to finish the LDSP, which was designed for children from 3;0-5;11. Parents of both children reported that the children had very good Chinese language skills. Language

[1] The identification numbers of the participants were coded as follows: "UC" stands for a child participant while "UA" stands for an adult. The numbers for the children, ranged from 1-75, were assigned in the order of the dates of testing. The adults were given the same number as their children. Therefore, UA22 was the mother of the child UC22.

samples collected did not demonstrate any language issues. The language scores for the other children were reported but should be interpreted with caution. The Chinese test was normed in Taiwan and the English test was standardized on monolingual English speaking children. Neither test was designed for children with the cultural and linguistic backgrounds in this study. Some test items (e.g., the picture of the soap, the truck, the construction site) in the LDSP were unfamiliar to the Chinese children growing up in the U.S. or China. Twelve participants who took the test fell beyond the target age range of the LDSP (i.e., 3;0-5;11). Thus, the two-year-olds were compared to the norms of three-year-olds and the six-year-old children were compared to the norms of five-year-olds. Although the two language tests were not ideal for the children in the present study, because no appropriate language tests were available, we used the two tests to gain a general measure of the children's language skills. The 42 children achieved a percentile rank of 21 to 93 for their total Chinese language scores. They all received much lower percentile ranks in their English total scores (range = 1-19), with a difference of 16 to 91 percentile ranks between the Chinese and English total scores (Appendix A).

Adults. Twelve mothers (age range = 27-45 years) of the child participants were recruited. All mothers indicated that Mandarin was their strongest and home language. One mother (UA56) had lived in Canada previously for seven years and had been in the U.S. for three months. The other mothers had been in the U.S. for 5 months to 10 years 3 months (mean = 6.2 years). One mother came from Taiwan (UA68) and the others came from China. Nine mothers reportedly started learning Mandarin from birth and three began learning Mandarin when they started school at five (UA22, UA47) and seven

(UA62) years of age. Six mothers (UA07, UA22, UA34, UA45, UA47, and UA62) were exposed to another Chinese dialect but spoke Mandarin as their native language.

Stimuli

Thirty familiar words (2 words x 15 disyllabic tone combinations) were chosen as the target stimuli based on two rounds of word familiarity testing (WFT1 and WFT2) conducted in three Chinese preschools in New York City. The children participated in the word familiarity testing did not participate in the present study. T3T3 combinations were excluded due to the T3 sandhi rule. There were 28 nouns and two verbs (<shua1ya2> 'brush teeth', and <he1shui3> 'drink water'). Hereafter, pinyin—the official Romanization system for Chinese characters—will be presented in angle brackets. The inclusion criteria for the words in the order of preference were: (1) high familiarity, (2) longer rime in the second syllable (S2), (3) less f0 perturbation and interruption in the initial consonant in S2, (4) longer rimes in the first syllable (S1), and (5) less f0 perturbation and interruption in the initial consonant in S1.

High familiarity was determined by the production rates (i.e., total number of children who produced the target word divided by total number of children who were asked to label the pictures in the two rounds of familiarity testing). For most of the DTs, the two words with the highest production rates in WFT1 and WFT2 were chosen. For T11 (T1+T1 combination), T12, T21, T23, another word with the next highest production rate but with a longer rime and/or voiced initial consonant in S2 and/or S1 was selected to replace one of the two words with the highest production rate (see reasons below). Eight words in the final stimuli were produced by less than 50% of the children in WFT1 and WFT2. Because no words with the same tone combinations had higher production rates,

these words were included in order to get a complete sample of disyllabic tone combinations. Appendix B shows the production rates of the 30 target words in the familiarity testing and by the children in this study. Most of the words (22 out of 30, 73%) were produced by more than 80% of the 44 children in this study. Only one word (<la1lian4> 'zipper') was produced by less than 50% of the children. (Appendix B).

The purpose of criteria (2) to (5) was to select words with longer voiced portions in the syllables of the target word. This allowed us to track as much as possible the f0 contours when doing acoustic analysis. More attention was given to S2 than to S1 because the greatest f0 variability occurs in S2, particularly in the initial portion of S2. Mandarin has mostly open syllables (i.e., with syllable final vowels) and the only final consonants are /n/ and /ŋ/. Words with longer rime were defined as words with complex vowels (e.g., diphthongs and triphthongs) and/or final nasals.

The goal of criteria (3) and (5) was to select words with less f0 perturbation and f0 interruption in the initial consonant of S1 and S2. Perturbations are the local raising or lowering of f0 following the consonant due to the change of intraoral and transglottal pressure during consonant production (Xu & Xu, 2004). Consonants that are produced with more oral obstruction (e.g., obstruent consonants) will cause an increase in intraoral pressure, which, thereby, decreases the degree of air pressure and air flow across the glottis, and, consequently, causes temporary changes in the f0. Consonants that are produced with minimal blockage of airflow (e.g., nasals and laterals), on the other hand, affect intraoral pressure minimally. When producing these consonants, the pressure at the glottis remains high and phonation continues with little (about one to two cycles of) f0 perturbations (Xu, 1999). Therefore, consonants that are produced with less airflow

obstructions were preferred. Because no glottal vibration of the vocal folds is involved in the production of voiceless consonants, which affects f0 tracking, voiced consonants are preferred to voiceless consonants. In cases of voiceless consonants, shorter consonants are preferred to longer consonants.

To meet criteria (3) and (5), the initial consonants in Mandarin were categorized into seven groups and ranked in the order from the lowest to the highest degree of f0 perturbation and interruption: (1) nasals (i.e., /m, n/) (2) approximants (i.e., /w, j, l, ɹ/) (3) unaspirated stops (e.g., /p, t, k/) (4) unaspirated affricates (/ts, tʂ, tɕ/ (5) fricatives /f, s, ʂ. ɕ, χ/ (6) aspirated stops /$p^h$, $t^h$, $k^h$/, and (7) aspirated affricates (/$ts^h$, $tʂ^h$, $tɕ^h$/) (See Appendix C). Initial consonants in the familiar words were prioritized and selected based on the above order.

The 30 colored drawings representing the 30 disyllabic words were duplicated to form two sets of pictures with different orders of presentation. Each participant was randomly assigned to one of the two presentation orders.

Data Collection

Collection of tone productions by the children and their mothers was conducted in the child's home, in the child's school, in a clinic or at the CUNY Graduate Center. All the procedures were administered in Mandarin except for the English language test for the children.

Child productions. Children attended one to two sessions. The sessions lasted from 30 minutes to two hours long, depending on the child's attention span and the number of breaks needed. Parents filled out a questionnaire in Chinese about the child's developmental backgrounds.

The child first completed the picture naming task and labeled the pictures two times in succession. Simple questions such as “这是什麽 [What is this]?”, “他在干吗 [what is s/he doing]?” were used to elicit productions. When the child failed to produce the target word, a toy object, a real object or gestures of the actions were presented. If the child still failed to produce the target word, semantic cues were given (e.g., “他很渴，他在做什么 [He is very thirsty. What is he doing]?”). In the 1320 trials (44 children x 30 words), children produced the target words in isolation at least one time in 1063 trials (80.5%), and in non-isolated positions (e.g, <da4 ping2guo3> ‘big apple’ instead of <ping2guo3> ‘apple’, <shua1ya2 ne> ‘brush teeth + sentence final particle’ instead of <shua1ya2> ‘brush teeth’) in 53 trials (4%). In 77 trials (5.8%), a non-target word was used to substitute for the target words (e.g., <yi1fu> ‘clothes’ for <la1lian4> zipper, <qing1wa1> ‘frog’ for <kong3long2> ‘dinosaur’). The target words were produced as monosyllabic words (e.g., <mian4> for <mian4tiao2> ‘noodles’, <dan4> for <ji1dan4> ‘egg’) in 25 trials (1.9%), as duplicated syllables (e.g., <mian4mian4> for <mian4tiao2> ‘noodles’, <bing2bing3> for <bing3gan1> ‘crackers’) in 24 trials (1.8%), and in English in 24 trials (1.6%). Children provided no response in 54 trials (4.1%) (Appendix D).

After the picture naming task, the Chinese language test was administered. Then a language sample was collected using picture telling, picture retelling, conversations and/or play. After that a hearing screening was performed. The English test was given at the end of the session so that the child would not be confused with the target language used. Children whose parents also participated in the study took a break and were then asked to imitate their mother’s productions (see below).

Adult productions. Mothers participated in the study were asked to fill out a questionnaire in Chinese about their language backgrounds. After the child had finished all the testing procedures described above, the mother was asked to label the same pictures two times in succession to the experimenter. Then she was asked to label the pictures one time to the child and let the child imitate her productions. These procedures were to elicit an adult-directed and child-directed register of speech. Child imitations and child-directed productions by the mothers were not analyzed in this study (see more details below).

All the children's and adults' productions were recorded on a digital recorder (Marantz, Model CDR420) in 16-bit PCM format at 44.1kHz sampling frequency through a Shure dynamic microphone (SM11).

## Perceptual Judgments of Children's and Adults' Tones

To determine children's accuracy in tone production, Mandarin-speaking adults were recruited to identify the children's and adults' tone productions.

### Stimuli

Natural/Unfiltered Stimuli. Forty-two nonsense natural (i.e., unfiltered) stimuli were used for two training blocks. The purpose of using nonsense words was to avoid lexical effects on tone judgment and to ensure that the judges could identify the tones without lexical knowledge. The first training block was composed of 12 (4 tones x 3 syllables) monosyllabic unfiltered nonsense words (e.g., <pou4>, <hen2>) and the second block consisted of 30 (15 tone combinations x 2 words) disyllabic unfiltered nonsense words (e.g., <pie2chuan4>, <diao3fo1>). All words had legal phonotactic constructions

and were recorded onto a computer in a sound treated booth by a Mandarin-speaking female speaker using Sound Forge 8.0 Software (Sony Media Software Inc., 2005). The 42 stimuli were normalized at -18.21dB, a high intensity level that did not cause clipping to the stimuli produced by the children and their mother. The tones of the 42 recorded unfiltered words were identified with 100% accuracy by a native speaker of Mandarin.

Filtered Stimuli. The 30 disyllabic nonsense words used in the second training block were low pass filtered at 400Hz using the Butterworth low-pass filter in Adobe Audition 2.0 ((Adobe Systems Incorporated, 2006) to form the third training block. The filtering procedure was to retain the f0 information and to eliminate most of the segmental information (Wong et al., 2005). Previous studies have reported that low-pass filter at 400 Hz was sufficient to eliminate most of the distinctive phonetic information while leaving prosodic information intact in adult speech (Cooper & Aslin, 1994; Friederici & Wessels, 1993; Jusczyk, Cutler, & Redanz, 1993). The purpose of the third training block was to familiarize the judges with the filtered stimuli like those used in the experimental blocks.

The stimuli for the experimental blocks were the target words produced by the children and the mothers in the picture naming task. Children's and mothers' first production of the target word was cut and saved as an individual sound file using Sound Forge 8.0 (Sony Media Software Inc., 2005). Only target words that were produced in isolation were adopted. Productions of non-target words, non-isolated productions, playful productions, tokens that were too loud (i.e., clipped), too soft (e.g., unintelligible mumbles) and noisy tokens (e.g., productions overlapped with another voice, tokens with clicks or pops or background noise) were excluded. If the first production could not be

used, the second production was selected. The 44 children contributed 999 usable spontaneous productions and the 12 mothers produced 353 usable adult-directed productions for tone judgment (see Appendices B and H).

We included 667 children's imitated DTs and 406 adults' child-directed DTs for tone judgment but they were not analyzed for the present study. The children's imitated productions were pilot data for a future study. Child-directed productions were excluded for three reasons. First, the quality of child-directed productions was not as good as adult directed productions because when we recorded the mother's child-directed productions and the child's imitations, the microphone was connected to the child to ensure better recording of the child's productions with the consideration that children often speak in low intensity. Second, only one child-directed production was recorded for each DT by each mother but there were more incidences of overlapping of the mother's and the child's voices in child directed productions. Third, some of the mothers produced fast speech instead of child directed speech. When the children saw the pictures, they tended to label them without waiting or listening to the mother's model because the children had labeled the pictures earlier in the session and listened to the mother saying the words to the experimenter and were, therefore, highly familiar with the pictures. As a result, some of the mothers tended to speak as fast as possible in an attempt to say the word earlier than the child or to avoid losing the child's interest in the task.

Adults' and children's productions were low pass-filtered at 400 Hz and 500 Hz, respectively, using the Butterworth low-pass filter in Adobe Audition 2.0 (Adobe Systems Incorporated, 2006), to eliminate the lexical information (Wong et al., 2005). Children's productions were filtered at a higher cut-off frequency because they tend to

have a higher fundamental frequency. All filtered stimuli were normalized at -18.21 dB and grouped by speakers.

There were 12 blocks of mothers' adult-directed productions (AD), 44 blocks of children's spontaneous productions (CP), 14 blocks of mother's child-directed productions (CD) and 27 blocks of children's imitated productions (CI). The 94 blocks of stimuli were arranged into 12 sets of experimental blocks. Each set included one block of AD, one block of CD (except for one set, which had 2 blocks of CD), three to four blocks of CP, and two to three blocks of CI. Effort was made to balance the age groups of the speakers in the CP blocks in each set of experimental blocks. There were one block of 2-year-olds', one to two blocks of three-year-olds', and one to two blocks of older children's productions (4- to 6-year-olds') in each set. The 12 sets of experimental blocks were grouped to form 6 pairs of experimental sets; each pair was designed for a judgment session for the judges. The age groups of the children who produced the CI blocks were semi-balanced in 6 pairs of experimental sets. There were four to five CI blocks in each pair of experimental sets, with one to three blocks from each of the age groups (2-year-olds, 3-year-olds, and children of four years and above). Overall, the number of trials (range = 403-406 trials), the number of AD, CD, CP and CI blocks and the number of DT productions from different age groups were comparable in the six pairs of experimental sets.

Judges

Five Mandarin-speaking judges between the age of 26 and 30 years were recruited. They all learned Mandarin from birth and reported Mandarin to be their dominant language. They were all doctoral students at the CUNY Graduate Center and had been in

the U.S. for eight months (J4 & J5) to nine years and six months (J2). The three male judges (J1, J3, J5) came from China and the 2 females judges (J2, J4) came from Taiwan. None of the judges had linguistic or phonetic training, or any difficulties with speech, language or hearing.

Procedures

Each judge attended two one-hour sessions each day, two days a week for two weeks. The two sessions on the same day were separated by at least an hour. All eight sessions were conducted in a quiet room at the CUNY Graduate Center. All sound stimuli were presented under headphones at a comfortable listening level by a customized computer program (Tagliaferri, 2005). The procedures of tone identification were first explained verbally to the judges, and again in writing on the computer screen. Judges were informed whether they would hear monosyllabic or disyllabic stimuli before the presentation of the stimuli. When the judge was ready, s/he clicked "start" on the screen. In the training session, the computer program then randomly presented a sound file in the training block. In the experimental sessions, the computer program randomly picked a block and presented a sound file in the block randomly. The judge could listen to the sound as many times as s/he wanted by clicking 'repeat' on the screen. S/he then indicated his/her decision by clicking the corresponding box indicating the tone (i.e., 1, 2, 3, 4) for the monosyllabic training block or tone combinations (e.g., 11, 12, 13, 14, 21…) for the disyllabic training blocks and the experimental blocks. After all the trials in a block were presented, the next block began. The judges were encouraged to take a break whenever they wanted and after each block. Their responses were automatically recorded on a spreadsheet.

In the first session, the judges listened to the three training blocks in sequence. Four of the judges made no errors in the first training block that involved 12 monosyllabic unfiltered nonsense words. J2 made two different tone errors and attained 83.3% accuracy. Four judges had accuracy rates of 90%-100% (0-3 errors) for the second training block that involved 30 unfiltered disyllabic nonsense words. J3 made seven errors and got an accuracy rate of 76.7%. All five judges performed very well in the third training block on 30 filtered disyllabic nonsense syllables. The accuracy rates were 93%, 100%, 90%, 97%, and 100% for the five judges, respectively. After the training blocks the judges filled out a language background questionnaire and had a hearing screening. They all passed the hearing screening of 500, 1K, 2K, and 4K Hz at 20 dB.

In the following six sessions (i.e., from the 2$^{nd}$ to the 7$^{th}$ sessions), the judges listened to a pair of experimental sets in each session in a random order. In the last session (i.e., the 8$^{th}$ session), the judges rerated the experimental blocks they listened to in the second session (the first session for experimental blocks) to establish intrajudge reliability. Because each judge had different sets of experimental blocks in the second session, the experimental blocks they rated again for intrajudge reliability were different.

## Results

Only adult-directed productions by the 12 mothers and spontaneous productions by the 44 children were analyzed. Because most of the data involved in the analyses violated the assumptions for parametric statistics (normal distributions and homogeneity of variance), we performed non-parametric statistics in all analyses, except for using Squared Pearson product-moment correlations ($r^2$) as a measure of effect size in examining developmental trends in DT accuracy across age groups.

First, interjudge and intrajudge reliability were examined. Second, the judges' accuracy rates on the adults' DT productions were investigated to determine whether the judges were able to identify the target DTs in filtered adult speech. Then data from adults and children were analyzed for the following purposes: (1) to determine whether the children's DT productions were adult-like, (2) to examine how children's DT accuracy changed with age, (3) to investigate the error patterns in children's DT productions, (4) to determine children's order of acquisition of the four Mandarin tones and context effects on children's tone accuracy rates, and (5) to investigate possible relations between children's demographic backgrounds and DT accuracy rates.

## Interjudge and Intrajuge Reliability

Categorizing DTs on filtered stimuli is an unfamiliar task for the judges and could be difficult for some native speakers of Mandarin. It involves extracting the tones in the auditory signal without lexical support, remembering the tone combinations, and selecting the right answer from 16 choices, which require attention, memory skills and meta-linguistic skills. Identifying children's DTs in filtered speech is even more challenging because children might produce f0 contours that did not fit any of the adult tone categories. The task could also be tedious to some judges because during the judgment sessions, the judges had to categorize about 3000 DT stimuli. In order to ensure that the judges were able to do the tasks reliably and consistently, vigorous analyses on the inter- and intra-judge reliability of the judges were performed.

Interjudge agreement was examined using both speakers and tones as sampling variables: (1) rank order of the 56 participants' (12 adults and 44 children) overall accuracy rates (summing across all DTs), (2) rank order of 44 child speakers on overall

accuracy rates, (3) rank order of the accuracy rates on the 15 DTs summing over all adult and child speakers and (4) rank order of accuracy rates on the 15 DTs summing over all child speakers. Kendall's Coefficient of Concordance was used to establish the overall agreement among the five judges. Spearman's rank-order correlations were computed to examine the agreement between each pair of judges.

The five judges (J1, J2, J3, J4 and J5) were highly intercorrelated on their rankings of the overall accuracy scores for the 56 speakers, [$W(N = 5, df = 55) = .920$], $\chi^2 = 253.106$, $p < .001$]; and for the 44 children, [$W(N = 5, df = 43) = .867$, $\chi^2 = 186.369$, $p < .001$]. Intercorrelations among judges on rankings of accuracy rates on the 15 DTs produced by the 56 speakers [$W(N = 5, df = 14) = .448$, $\chi^2 = 31.380$, $p = .005$]; and the accuracy rates of the 15 DTs produced by the 44 child speakers [$W(N = 5, df = 14) = 0.545$, $\chi^2 = 38.119$, $p < .001$] were considerably lower but still significant. Spearman's rank-order correlation coefficients of each pair of judges showed high correlations, with $r_s$ values ranging from .863 to .936 for the overall accuracy scores (combining across the 15 DTs) of the 56 speakers and from .764 to .895 for the 44 child speakers (Appendix E). However, interjudge correlations on the 15 DTs were much lower. Spearman's rank-order correlation coefficients ranged from .082 to .800 for the 15 DTs produced by the 56 adult and children, and from .068 to .821 for the 15 DTs produced by the child speakers. Only three judges (J2, J4 & J5) correlated significantly with one another on the accuracy rates of the 15 DTs produced by the 56 speakers and 44 child speakers (Appendix E). J1 and J3 were, therefore, excluded in further analyses.

With the exclusion of J1 and J3, even higher overall interjudge reliability was obtained. High intercorrelations were found among the three judges on overall accuracy

rates for all 56 speakers (W = .941, p < .001), and for the 44 child speakers (W = .900, p < .001), with slightly lower overall correlations on rankings of the 15 DTs for all 56 speakers (W = .854, p = .001) and for the 44 children (W = .860, p = .001). Spearman correlations of each pair of judges were also quite high on overall accuracy rates for all 56 speakers ($r_s$ ranged from .893 to .936) and for the 44 children ($r_s$ ranged from .831 to .889), but were somewhat lower for the 15 DTs produced by all 56 speakers ($r_s$ ranged from .743 to .800) and by the 44 child speakers ($r_s$ ranged from .750 to .821) (Appendix E).

Spearman's Rank-Order correlation coefficients were computed to examine the test-retest reliability of each of the three selected judges (J2, J4, & J5) on four measures: (1) the rank order of the overall scores of all the speakers (adult + children), (2) the rank order of the overall scores of the child speakers, (3) the rank order of the accuracy rates of the 15 DTs produced by all speakers, and (4) the rank order of the accuracy rates of the 15 DTs produced by the child speakers in the test and retest sessions.

J2 demonstrated very strong test-retest reliability in all four measures ($r_s$ ranged from .914 to .967). J5 showed strong test-retest reliability on the overall accuracy on all speakers and child speakers only ($r_s$ = .946 and .893, respectively). Test-retest reliability on the 15 tones produced by all speakers and by the child speakers was slightly lower ($r_s$ = .733 and .732, respectively). J4 had high intrajudge reliability when using speakers as the sampling variable ($r_s$ = .912 and .833 for the accuracy on adult and child speakers and child speakers, respectively). However, she showed low test-retest relations on the accuracy of the 15 tones produced by all the speakers ($r_s$ = .338) and by child speakers ($r_s$ = .286) (Appendix F).

To investigate the contributing factors to J4's reduced intrajudge reliability, her performance in the test and retest sessions was compared. In both sessions J4 categorized the 15 DTs produced by adults with very high reliability. Only one T11 word produced by UA68 was categorized differently in the test and retest sessions. Therefore, the contributing factors of J4's reduced intrajudge reliability on the accuracy scores of the 15 DTs were from the 15 DTs produced by the children. The number of different categorizations, which was defined as the number of child productions for which different DTs were chosen by the judge in the test and retest sessions, was compared for each of the 15 DTs. The results suggested that J4's differences in her categorization of T11 and T43 in the test and retest sessions were the main factors for her low intrajudge reliability. J4 had the most number of different categorizations on T43. All the eight different categorizations for T43 involved selecting a wrong DT in the test session but the correct DT in the retest session. On the other hand, all the five different categorizations J4 made for T11 involved categorizing the productions correctly in the test session but incorrectly in the retest session, a reversed categorization pattern that was not observed in J2 or J5. The higher number of difference categorizations for T43 and the direction of categorization differences for T11 could have contributed to very different rankings of the accuracy rates of the 15 DTs in J4's test and retest sessions. Given that J4 was highly reliable in categorizing adults' DTs, her differences in categorizing children's DTs in the test and retest sessions might be due to a shift of using different cues for tone categorization when the DTs were not in adult forms.

To examine whether J4 categorized T11 or T43 more differently than J2 and J5, the number of correct and incorrect judgments of the DTs by all speakers in the

experimental sessions were compared among the three judges. J4 did not make more judgment errors than the other two judges in any of the 15 DTs.

Despite J4's lower intrajudge reliability on the 15 DTs, we included her in our analysis because she demonstrated high interjudge reliability with J2 and J5 on the overall scores of the speakers and the accuracy rates of the 15 DTs produced by all the speakers and the child speakers (see interjudge reliability above). Her accuracy rates on the 15 DTs produced by the 44 children and by the 12 adults were comparable to those of the other two judges. She did not make more errors than the other two judges in any of the 15 DTs produced by the children or by the adults. All subsequent analyses presented below were based exclusively on the judgments of J2, J4 and J5.

## Judged Accuracy of Adults' Disyllabic Tone Productions

The term accuracy is defined as the judges' correct identification of the intended target tone combination produced by the speaker. The judges identified the adult speakers' DTs with high accuracy. All adult speakers' DT productions were identified with over 95% accuracy (range = 95.4%-98.9%), except for UA42 (overall accuracy rate = 87.4%. T41, T43, T44 were identified with the lowest accuracy rate at 50%). Of the 353 usable tokens produced by the 12 mothers, 323 productions (92%) were correctly identified by all three judges. Among the 30 productions that were not correctly categorized by all judges, 21 productions (5.95%) were judged incorrectly by one judge, five (1.42%) by two judges and only four (1.13%) of them were incorrectly identified by all three judges. Seven of the 30 erroneous productions involved T11, six involved T43, and five involved T14 (see Appendix G). None of the adults produced any DT that was

categorized with lower than 50% accuracy. Overall, the judges were able to identify the target tones accurately from the filtered stimuli under the task demands of the experiment.

## Comparison of Adults' and Children's Tone Production Accuracy

Four major comparisons were made to determine whether children's DTs were adult-like. First, the 44 children's overall accuracy rates summing across the 15 DTs were compared to adults' to investigate whether children as a group produced the DTs in an adult-like manner. Then the children were divided into 4 different age groups and their accuracy, summing across the 15 DTs was compared to adults' to determine if any of the age groups produced the DTs in an adult-like manner. After that, children's accuracy rates on each DT were compared to adults' to examine whether children as a group produced any of the 15 DTs in an adult manner. Finally, children's accuracy rates on each of the 15 DTs were compared with adult performance by age groups to determine which DTs were produced with adult-like accuracy in each of the age groups.

### Overall DT Accuracy

Children as a Group vs. Adults. As a first comparison, the 44 children's correct production scores collapsing over all 15 DTs were compared with the adult group. A Mann-Whitney U test of independent group differences indicated that children's accuracy rates (mean rank = 22.68) were significantly lower than adults' (mean rank = 49.83), $U(N = 56) = 8.000$, $p < .001$; that is, filtered versions of children's productions were not judged as the intended tone sequences as often as were the adults' productions.

Note that the above analysis collapsed over cases in which only one word was produced for a particular DT and cases in which both words were produced. In the case

where only one word was produced, the accuracy rate for the DT was based on 3 judgments (1 word x 3 judges); thus, 100% accuracy represented three correct judgments out of three judgment trials. In the case where both words for the same DT were produced, the accuracy rate was determined by six judgments (2 words x 3 judges); therefore, 100% correct denoted six correct judgments out of six trials.

Younger children produced fewer productions of both words for each DT. Table 2 shows the number and percent of two-word, one-word and no productions for the DTs by age group. Total number of cases was equal to 15 DTs multiplied by the number of speakers in the age group. The number of two-word productions was defined as the number of times a child labeled both pictures for the same DT with the target words. Two-year-old (C2), three-year-old (C3), four-year-old (C4), five- and six-year-old (C5+) children and adults produced both words for a DT in 53%, 66%, 78%, 90%, and 100% of the cases (See the last column in Table 2). Among the productions that were included for tone judgment, there were more one-word DT productions by C2 (70 productions, 38.9%) and C3 (67 productions, 34.4%) than by older children (43 productions, 26% for C4; 21 productions, 8% for C5+).

In view of the different proportions of one- and two-word production rates for the DTs produced by each age group, analyses based on only one of the two words for each DT were performed such that the accuracy rates of the DTs was consistently based on three judgments. The 30 target words were divided into two groups based on the number of child productions included in this study. Between the two words for each DT, the ones with more usable tokens (i.e., number of child productions included in this study) were categorized as High Words and those that had fewer usable tokens were categorized as

Low Words (Appendix B). The two words for T23 had the same number of usable child productions in this study; the one that had higher production rates in the word familiarization tests was selected as the High Word. All the 15 High Words were also words with higher production rates in the familiarization tests and by the children in this study, except for T21 and T24. These two DTs had comparable production rates in the familiarization tests; the word chosen as High Word had a production rate of only 2%-5% lower than that of the other word (see Appendix B). Altogether there were 176 and 552 High Words and 177 and 447 Low Words produced by adults and children, respectively. Appendix H presents the number of High and Low Words produced by different age groups.

The children's and adults' overall accuracy rates summing across the DTs in the 15 High Words were compared. The results were similar to those found with all 30 words included: children as a group did not produce the DTs in High Words as accurately as adults, $U (N = 56) = 4.5$, $p = < .001$.

Children in Different Age Groups vs. Adults. Because there was large variability in the overall accuracy rates of the DTs among the children across the age range, children were divided into four age groups (C2, C3, C4 and C5+) and their overall accuracy rates were compared to adults' to determine if any of the age groups was adult-like. Figure 3A shows boxplots of overall accuracy rates, summing over all DTs and all 30 words for these age groups in comparison with adults. Figure 3B shows the results for the High Word analysis. Even the five-year-old children's tone accuracy rates were significantly different from adults' for comparisons based on 30 words [$U (N = 20) = 6.000$, $p = .001$], and High Words [$U (N = 20) = 4.500$, $p = .001$]. The results, thus far, consistently

suggested that children in the four age groups did not produce the DTs as accurately as adults. Note that median performance for the C3 and C4 groups show a decrease in performance across this age range. This "reverse" in developmental trend is discussed further below.

Accuracy on Individual DT Combinations

Although children's overall accuracy rates summing across the 15 DTs were not adult-like, it remained unclear whether some of the DT combinations were produced with adult-like accuracy. Thus, children's accuracy rates on each of the 15 DTs were compared to adults', first as a single group, then by the four age groups.

Children as a Group vs. Adults. Figure 4 shows the boxplots of the 44 children's and the 12 adults' correct production scores for each of the 15 DTs based on their productions of the 30 words (Figure 4A) and the 15 High Words (Figure 4B). Most of the adult scores were at ceiling. Children as a group showed substantial variability in most of the DTs compared to adults. For each of the 15 DTs, there were children whose accuracy rates fell within the adults' distribution. On the other hand, for most of the DTs, there were children who performed at or near the floor. Results of Mann-Whitney U test revealed that children produced most of the 15 DTs (except T14 in All Words and T11 and T14 in High Words) with accuracy rates significantly different from adults', indicating that most of the DTs produced by the children were not adult-like (see Appendix I), although median accuracy rates for some DTs (T14, T21 T32 when all words were included, and T14, T21, T23, T24, T32 and T43 when only High Words were included) were at ceiling for the children.

Children in Different Age Groups vs. Adults. Next, children's accuracy on the 15 DTs was compared to adults' by age group. Figure 5 shows the accuracy rates of the 15 DTs on all 30 words (Figure 5A) and on the 15 High Words (Figure 5B) by age group. The error bars represent two standard errors around the mean scores. The dotted line marks the lower bound of the 95% confidence interval of the distribution of the adults' scores for the DT. Some DTs (5 DTs for All Words and 10 for High Words) were produced with 100% accuracy by all adults. For these DTs, the highest value (95.55%) of the lower bound of the 95% confidence interval among the other DTs was adopted and marked with a solid line in the charts. The adults' productions for the 15 DTs were mostly at ceiling with little variability (Figures 5A and 5B). There were age group differences on the amount of overlap of the adults' and children's score distributions. Distributions of C5+ children's scores overlapped with adults' for most DTs. Most C2 children's distributions did not overlap with adults'. Again, when the High Words only were included (Fig 5B), accuracy rates for C3 and C4 groups showed a reversal in the developmental trend for nine of the 15 DTs [10 when all 30 words were included (Fig 5A)].

Table 3 lists the lower bound of adults' 95% confidence interval for each DT (again, for the DTs which all adults produced with 100% accuracy, the 95% confidence interval was set at 95.6%) and the percentage of children in each age group whose accuracy rates on the DTs were comparable to adults' (i.e., with accuracy rates equal to or higher than the lower bound of the 95% confidence interval of the adults' score distributions) in All Words and in High Words (See Appendix J for details). In general, the percent of children who produced the DTs with adult-like accuracy rates increased

with age. This pattern was more prominent in High Words when all speakers were judged on the same word for the same DT.

As the table shows, very few DTs (T32 in All Words and T24, T31, T32in High Words) were produced with adult-like accuracy by a majority of the C2 children (Table 3). If "mastery" of a DT before age 3 years is defined as a majority of C2 children showing adult-like accuracy, two DTs (T32, T24) were mastered the earliest (75% and 50% for All words, respectively, and 75% and 71% for High words respectively), T12 was the least well mastered by 2-year-olds. No C2 children produced this DT with adult-like accuracy in either All Words or High Words analyses.

Three-year-old children produced more DTs with adult-like accuracy (Table 3). Four (T11, T14, T21, T43) and seven (T11, T14, T21, T23, T32, T41, T43,) of the 15 DTs in All Words and in High Words, respectively, were mastered by the majority of the children in this age group. T14 was mastered by the most C3 children (67% and 83% for All Words and High Words, respectively), and T21 was the second most mastered DT (58% and 67% for All and High Words, respectively). T12 and T22 were mastered by the fewest three-year-old children (approximately 20% and 30% for All Words and High Words, analyses respectively).

The number of four-year-old children who produced the DTs with adult-like accuracy was not higher than three-year-old children (Tables 3). Four (T11, T21, T32, T41,) and five (T11, T21, T23, T24, T32, ) DTs in All Words and High Words, respectively, were produced with adult-like accuracy by a majority of C4 children. Eleven DTs in All Words and 9 DTs in High Words were mastered by fewer C4 than C3 children using this criterion. Only T11, T22, T23, T32 in All Words and T11, T13, T21,

T22, T24, T32 in High words were mastered by more C4 than C3 children. T32 was mastered by the most C4 children (70%). T21 was produced next best (55% and 82% for All Words and High Words analyses, respectively), as well as T11 (64% in both All Words and High Words analyses). T12 and T42 were mastered least well by four-year-olds (0% and 20% for All Words, 0% and 20% for High Words).

The rate of DT mastery according to this criterion shows a large increase for C5+ children for most DTs (Table 3). Ten and 13 of the DTs in All Words and in High Words, respectively, were produced with an adult-like accuracy rate by a majority of the children in this age group. T11, T14, T21 were the easiest and were produced with adult-like accuracy by 100% of the children in the group. T12 and T24 remained difficult for the majority of these children when both High and Low words were analyzed (Table 3).

The DTs mastered across the four age groups all involved DTs with compatible (C) f0 contours (i.e., the offset of the f0 contour for the tone in S1 is close to the onset of the f0 for the tone in S2). On the other hand, three of the DTs least well mastered (T12, T22, T31) involved non-compatible (NC) f0 contours (i.e., the offset of the f0 contour for the tone in S1 was very different from the onset of the f0 for the tone in S2). There are eight NC DT combinations: T12, T13, T22, T23, T31, T34, T41, and T44 (see Table 1). If the DTs in each age group were ordered in descending order by the percentage of children who produced the DTs with adult-accuracy, five to seven of the eight NC DTs fell in the bottom half (8) of the lists (i.e., produced with adult-like accuracy by the smallest percentage of children) in all age groups for All Words and High Words, with an exception in the High Words produced by C2 children. Only half (4) of the NC DTs fell into the bottom half of this list. The effects of tone compatibility on children's accuracy

rates are discussed further in the sections on error patterns and order of accuracy rates of the four Mandarin tones below.

In summary, these results indicated that some DTs were produced by more children with adult-like accuracy than others. The number of children who produced the DTs with adult-like accuracy increased with age. C4 children showed some regression in DT production mastery. Overall, there was a tendency for children to master DTs with compatible tone contours earlier than DTs with non compatible tone contours. The developmental trends shown here are discussed further in the next section.

## Developmental Trends in Children's DT Productions

### Development of Overall DT Accuracy

Children as a Group. As a first step, children's overall accuracy, summing across the 15 DTs, was subjected to correlational analysis. In this comparison, percent of judged correct productions was computed over the usable words produced by each child. Figure 6 presents scatterplots of children's overall accuracy rates on the 30 target words (Figure 6A) and on the High Words (Figure 6B) as a function of age. There was a significant positive relation between age and percent correctly produced DTs on the 30 words [$r_s(N = 44) = .443$, $p = .003$] and on High Words [$r_s(N = 44) = .453$, $p = .002$]. The $R^2$ of .199 for All Words and .208 for High Words indicated that age accounted for approximately 20% of the variability in the children's performance on DT production. As the scatterplots show, children between 2 and 4 years old varied greatly in their ability to produce DTs accurately, as judged by native Mandarin listeners. Accuracy of children in the same age group varied substantially in producing the 30 words, with a difference of about 50% between the highest and lowest scores in C2 (range = 27%-78%), C3 (range =

40%-88%) and C4 (range = 38%-87%). Although the number of 5 - 6 year olds tested was rather small (n = 8), their scores were much less variable (range = 67%-94%), reflecting mastery of many more DTs (as discussed above).

When all 30 words were taken into consideration (Figure 6A), most C2 children's scores clustered below 70% accuracy (10 out of 12 children), with only two older C2 children producing the DTs with overall accuracy rates above 70%. Using C2 children's distribution of overall performance as a reference, 7 of 13 three-year-old children achieved overall accuracy rates over 70%. Younger and older children in C4 seemed to show a somewhat different pattern of performance in the All Words data. Three of 5 younger C4 children produced DTs with over 70% overall accuracy, whereas only one of six older four-year-olds had accuracy rates above 70%. Most of them had lower accuracy rates than their younger peers. Six out of eight C5+ children had overall accuracy rates of over 70%; 3 achieved accuracy rates over 90% (range = 91%-94%); and the overall score of one child fell within the 95% confidence interval of the adults' scores on the 30 words (Figure 6A).

Because the mastery data for the All Word analysis presented above suggested possible age group differences on DT accuracy rates, children were again grouped into four age groups to analyze age trends (Figure 3A). Kruskal-Wallis analysis of variance was conducted to examine whether there were any age group differences on the overall accuracy rates of DTs. Significant differences were found among the age groups, $\chi^2$ (3, N = 44) = 12.270, p = .007. Post hoc Mann-Whitney tests were used to compare the percent correct scores across the age groups. A Bonferroni correction was applied and corrected p-values of .0083 and .0017 were adopted for the alpha levels of .05 and .01, respectively.

The mean rank of children's scores was significantly higher for C5+ children (15.69) than for C2 children (7.04), ($U = 6.500$, $p = 0.001$), indicating that children who were over 5 years old produced the DTs significantly better than 2-year-old children. No significant differences were found between other age groups at 0.05 or 0.01 level with adjusted p-values (see Figure 3A).

Similar developmental trends with slightly better performance by older C4 children were observed among the age groups in the overall accuracy rates of the High Words (see Figure 6B). Two children in C2 achieved over 70% accuracy. A reverse in development of accuracy of C3 and C4 children continued to be evident. The two oldest children in C4 had higher overall accuracy rates (an increase of 11% for both children) on High Words than on all the 30 words, which reduced the differences in the accuracy rates between younger and older C4 children. All eight 5-year-olds had accuracy rates over 70%; three had over 90% accuracy. One of these scores was within the 95% confidence interval of the adult accuracy rates.

Mann-Whitney U tests were conducted to analyze significance of developmental trends among the age groups (see Figure 3B). Results of pairwise comparisons confirmed that children in C5+ had significantly higher overall accuracy rates on DTs than two-years-old [$U(N = 20) = 7.000$, $z = -3.166$, $p = .002$*] and four-year-olds [$U(N = 19) = 8.5$, $z = -2.934$, $p = 0.003$*], but not three-year-olds [$U(N = 21) = 18.000$, $z = -2.464$, $p = .014$] after the p-values were adjusted for multiple comparisons (p-values of .008 or .002 were required for significance at 0.05 and 0.01 level, respectively).

In summary, although considerable variability was observed in the accuracy rates of children in the same age group, children's overall accuracy rates of DTs improved

with age. Four-year-old children appeared to produce DTs less accurately overall than three-year-old children. Five- and six-year-old children's DT productions were significantly better than 2-year-olds and the variability in children over 5 years of age decreased considerably.

Developmental Trends for Individual DT Combinations

While overall accuracy rates showed that children's performance improved with age, the results of individual DT contours in the previous section with respect to mastery suggests that the developmental trends differed across the 15 DTs. Performance on individual DT combinations was inspected further using the overall scores for the four age groups. Figure 7A presents performance on both words, with an overall regression line indicating the rate of development over age groups for each DT. The accuracy rates for the High Words and Low Words were marked by circles and triangles, respectively. Figure 7B presents data on the most frequently produced word for each DT (High Words). Positive correlations between accuracy rates and age were found for all 15 DTs in All Words and High Word comparisons. However, the large variations (range = .004 - .943) in the coefficient of determination ($R^2$) of the 15 DTs suggested that children developed different DTs at different rates, and that specific lexical effects altered the overall trends for some DTs.

Two-year-old children demonstrated the most lexical effects in terms of the number of DTs that were produced with more than a 20% difference in the High and Low Words and the amount of difference in the accuracy rates for the High versus the Low Words (see the gaps between the High (circles) and Low (triangles) words in C2 in Figure 7A). Part of the reason for the lexical effects in the DT accuracy rates for C2

children may be due to limited data for the Low Words. There were only three to four productions for T14, T24, and T42 in the Low Words by C2 children. These DTs had big lexical effects (38%-42% difference between High and Low Words) at C2.

Different speakers producing the High versus the Low Words could be another contributing factor to the large lexical effects in some of the DTs by C2. Only 2 two-year-old children produced both the High and Low Words for T14, T23, T24, T42 (see Appendix K), which were four of the seven DTs that showed a more than 20% difference in the accuracy rates of the High and Low Words by 2-year-old children (range = 38% to 48%). Thus, the High and Low Words for these DTs were produced mostly by two different groups of C2 children. Given that young children had great variability in their DT productions, the overall DT accuracy rates may have been affected by limited productions and productions from different young children. All DTs produced by the same eight or more C2 children for both the High and Low words (T11, T31, T34, T43, T44) showed relatively smaller lexical effects (accuracy rate difference < 10%, except for T44 which had an 18% difference in the accuracy rates).

Lexical effects (accuracy rate difference > 20% in High versus Low Words) in DT productions were also observed in older children but to a reduced degree; none of the differences in the accuracy rates between the High and Low words for the same DT produced by children three years or older exceeded 40%. Except for T14 by C3 children (accuracy difference = 28%), for which the Low Word was produced by 3 children only, all other DTs (T34 in C3; T12, T21, T22, T41 in C4; and T12, T22, T31, T42, T43 in C5) were produced by five or more children in the age group and had 5 or more speakers who produced both the High and Low Words for the DTs. Thus, for these DTs, the

contributing factors for the lexical effects were less likely due to limited data points or productions of the High and Low Words by different speakers.

To reduce these lexical effects and to use the most reliable data for the youngest age group, effect sizes computed across age groups for the High Words only analysis were used for comparing growth rates of individual DTs. Table 4 shows the accuracy for each DT by C2 and C5+ children, the coefficient of determination ($R^2$) as a measure of effect size, and a verbal descriptor of effect size on High Words. Given the wide range in $R^2$ values, they were divided into five categories of effect sizes: none ($R^2 < .1$), small ($.1 \leq R^2 < .3$), medium ($.3 \leq R^2 < .5$), large ($.5 \leq R^2 < .75$), and very large ($R^2 \geq .75$). On the left, the DTs are arranged by the tone in Syllable 1, whereas on the right, the same data are rearranged by Syllable 2 to promote easier comparison of tone combinations.

DTs showing no or small growth rates included those that even the youngest children produced relatively accurately (accuracy rates ranged from 54% to 83%): T32, T31, T24, T42, T44 (Table 4). However, three other DTs (T13, T21, T34) with accuracy rates of over 50% by C2 children showed medium to large effect sizes, indicating more rapid development from 2 to 5 years. Although DTs that were produced with higher accuracy rates at C2 were expected to have reduced growth functions due to relatively more limited room for growth, it was noted that most of these easier DTs at C2 were not produced with the highest accuracy by C5 children. DTs that had accuracy rates below 60% at C2 demonstrated a large variety of growth functions, ranging from small to very large. As expected, there was a general tendency for more difficult DTs (those with the lowest accuracy rates at C2) to have large or very large growth rates. T12

appeared to be the most difficult DT across age. It was produced with the lowest accuracy by C2 children, underwent a medium rate of change and remained the most difficult DT for C5 children (57%). Finally, T24 showed a slight decrease in accuracy over age.

In general DTs starting with T3 appeared to show the least change with age (see Figure 7B row 3), whereas DTs ending with T3 showed large to very large increases in accuracy with age (see Figure 3B column 3). Similar context effects are apparent for the other tones. Thus, trends in DT development appeared to be influenced greatly by the combination of tones present in these disyllable words.

## Error Patterns

This section focuses on analyzing children's misidentified DT productions to discover the error patterns. Children's incorrect productions were examined from five perspectives: (1) error patterns in terms of number and consistency of incorrect judgments for each DT production, (2) number of one- versus two-syllable errors, (3) number of errors in S1 versus S2, (4) number of errors in C versus NC tone combinations, and (5) major substitution patterns for the misidentified DTs.

### Number and Consistency of Incorrect Judgments of DT Productions

DT productions by different age groups tended to be misidentified by different numbers of judges. Table 5 shows the proportion of children's and adults' DT productions in High Words that were misidentified by one or more judges. Children in C2-C4 groups produced more DTs that were misidentified by more judges than C5+ children. In all, 30%-47% of the productions by C2-C4 children were miscategorized by

2 or 3 judges (sum of rows 2 and 3 in Table 5), while most of the errors by C5+ were incorrectly identified by only one judge. Approximately 20%-30% of the productions by C2 to C4 children, but only 7% of the productions by C5+ children, were judged incorrectly by all 3 judges. Among the DTs that were misidentified by all 3 judges, half of the time they were categorized as the same (incorrect) DT by all three judges. The fact that younger children's DT productions were misperceived by more listeners and many DTs were categorized into different DTs by different listeners suggests that younger children might have produced DTs with f0 contours that were very different from the target forms of the four Mandarin tones. When the listeners were forced to identify these tones, they might select from two or more equally possible choices, select a tone when none of the four tones was appropriate, or give more weight to certain cues in the f0 contours to decide on the target tones.

Number of Tone Errors in Children's Disyllabic Productions

When children's productions were judged as different from the intended DT, the judged DT could differ from one or both intended tones. Figure 8 shows the boxplots for tone errors, summed over all children's productions that consisted of errors in one syllable (DTs with errors in S1 only, in S2 only, or in either S1 or S2) versus errors in both syllables. Results of Wilcoxon Signed Ranks tests revealed that children made significantly more errors in one syllable (either S1 or S2) than in both syllables. Number of DTs with errors in only S1 ($N = 44$, $z = -4.647$, $p < .001$), in only S2 ($N = 44$, $z = -2.798$, $p = .005$), and in either S1 or S2 ($N = 44$, $z = -5.712$, $p < .001$) were all significantly greater than errors in both syllables. The results indicated that children tended to produce at least one of the intended tones correctly, and infrequently made

errors in both syllables of the DT. This suggested that children focused their attention on one of the syllables when attempting to produce the disyllables. In the next section, error patterns as a function of syllable position and compatibility of tones are examined.

Number of Tone Errors in Syllable 1 versus Syllable 2

As presented in the introduction, in adult productions, the f0 contour of a particular tone can be more complex in S2 than in S1 because of carryover coarticulatory influences (Xu, 2001). In adult productions, the f0 contour of a specified tone in S1 changes relatively little as a function of context (see Figure 2), whereas the f0 contour in S2 involves a transitional portion from the offset of the tone in S1 such that the canonical tone pattern in S2 is realized in the latter part of the syllable. Thus, it was predicted that children would produce tones less accurately in S2 than in S1, particularly in non-compatible (NC) contexts where the coarticulatory variation is greatest. A Wilcoxon Signed Ranks test was used to determine whether children made more errors in S2 than in S1. Counter to our prediction, children overall made significantly more errors overall in S1 than in S2 ($N = 44$, $z = -1.962$, $p = .050$). Most of the age groups showed this error pattern; 6 (of 12), 9 (of 13), 10 (of 11), and 6 (of 8) children in C2, C3, C4 and C5+ groups, respectively, had lower accuracy rates in S1. Only C4 children's differences were statistically significant ($N = 11$, $z = -2.805$, $p = .005$), probably due to small sample sizes which reduced statistical power of tests of age groups.

Number of Tone Errors in Compatible versus Non-compatible DT Combinations

Due to larger differences between the f0 at the end of the first tone and at the beginning of the second tone in NC tone combinations, more complex f0 contours in

terms of the changes in the direction and speed are observed in adult productions of NC DTs. If children's accuracy rates were related to f0 complexity, then, we hypothesized that they would have more difficulties with NC than C DTs. Wilcoxon Signed Ranks tests confirmed that children produced C DTs significantly better than NC DTs, N = 44, z = -3.169, p = .002. Eight (of 12), 9 (of 13), 7 (of 11) and 6 (of 8) children in the C2, C3, C4 and C5+ groups, respectively, showed the same pattern. However, only C4 children reached significance level, N = 11, z = -1.956, p = .050, again, probably due to the small sample sizes.

Further analysis was performed to examine the interaction of syllable position and the compatibility of tones on children's DT production accuracy. Children's overall tone accuracy rates, summing across the four tones, were computed for four conditions: (1) in S1 followed by a C tone in S2, (2) in S1 followed by a NC tone in S2, (3) in S2 preceded by a C tone in S1, and (4) in S2 preceded by a NC tone in S1. Results revealed that, in S1, children produced the tones in C contexts more accurately than in NC contexts (N = 44, z = -4.450, p = .000). Though C2 (9 of 12) and C5 (7 of 8) children also showed the same pattern, only C3 and C4 children reached statistical significance, N = 13, z = -2.703, p = .007 for C3 children, and N = 11, z = -2.847, p = .004 for C4 children. There were no significant differences in the accuracy among the tones in C versus NC contexts in S2 and in C contexts in S1.

Given that significant differences were found in the accuracy rates of tones in C vs. NC contexts in S1, further analysis was carried out to examine whether the same pattern was observed for all the four tones in S1. For each of the four tones in S1, their accuracy rates preceding a C tone versus preceding a NC were compared. The results

showed that T2, T3 and T4 in S1 were produced significantly more accurately in C than in NC contexts ($z = -2.236$, $N = 37$, $p = .025$; $z = -2.878$, $N = 38$, $p = .004$; $z = -2.647$, $N = 42$, $p = .008$ for T2, T3, and T4, respectively). Due to missing data in some of the conditions for some children, no age group comparisons were performed.

Overall the results showed that children's accuracy rates were related to f0 complexity. They had more difficulties with NC than C tone combinations. The fact that children made more errors in NC contexts than in C contexts in S1 suggested that when children produced NC tone sequences, they tended to modify the f0 contour in S1. If children modify the f0 in S1 when producing NC DTs such that the transition from S1 to S2 is less abrupt in S2 (i.e., making the f0 contour in S1 more compatible to that of S2), the f0 complexity in S2 in NC DTs would be reduced and, therefore, no significant difference would be found in C vs. NC contexts in S2. Also, if the production of NC DTs mostly involved a change of f0 in S1 or having the f0 transition in S1 rather than in S2, children would have more tone errors in S1 than in S2. Evidence that supports this speculation is presented in the following section on major substitution patterns and in the section on acoustic analysis.

Major Substitution Patterns for Children's Misidentified DTs

This section examined how children's incorrect DT productions were perceived by listeners. Table 6 shows the judges' responses to the DTs produced by two- to four-year-old children. Correct identification (in percent) of the target DTs are shown on the diagonal highlighted in black. All other cells ($n = 210$) represent potential substitution errors (use of another DT to substitute for the intended DT) for the target DTs. Note that C5+ children were not included because they made significantly fewer errors than

younger children (compare Appendix Q to Appendices N-P). Without including C5+ children, more major substitution patterns representing the major errors of C2-C4 children were found and these patterns also covered most of the major substitution patterns of C5+ children (compare Table 6 and Appendices M-Q). As shown in Table 6, the judges identified a wide variety of DTs (n = 120) as substitutes for the intended target DTs. The number of substitution patterns used by the judges decreased with the age of the speakers. Altogether 78, 58, 64, 25 and 9 substitution patterns were used to categorize the DT productions of C2, C3, C4, C5+, and adults, respectively (Appendices L and N-Q).

Some substitution patterns were perceived more often than others. Seventeen major substitution patterns (substitution patterns that accounted for more than 10% of the total judgments for the target DT) were found in the confusion matrix for C2-C4 children (highlighted in grey in Table 6). For easier comparisons they are also listed in Appendix R. These major substitution patterns were seen for intended DTs that constituted seven NC patterns (T12, T22, T23, T31, T34, T41, T44) and five C DTs (T11, T14, T24, T42, T43). Note that the percentage presented in each cell was computed by counting the number of times the DT was chosen by any judge for the target DT produced by any children and divided the number by the total number of judgments made for that target DT. Thus, the percentage might not represent how consistently a child's production was identified by the judges. For examples, in Table 6, T41 and T43 were used by the judges to substitute for T42 comparably (12% and 13%, respectively). However, no children's T42 production was consistently heard as T43 by all three judges, whereas three child productions for T42 were heard by all three judges as T41. In any case the substitution

patterns in the confusion matrixes reflected the perceptual errors the judges made on the children's DT productions.

Eleven of the 17 major substitution patterns for C2 to C4 children's DTs had an incorrect tone in S1 (Table 6 and Appendix R). Six of them involved substituting the target tone in S1 with another tone that had an f0 at the end of the syllable closer to the beginning f0 for the tone in S2. For example, in T31→T21 (T21 was used to substitute for T31) and T34→T24, the intended tone in S1 (T3) was a low tone and ended with a low f0. The intended tones in S2 (T1 and T4) started with a high f0. However, when children produced these DTs, the tone in S1 was heard as T2, a tone that ended with a high f0, closer to the beginning f0 of T1 and T4 in S2. These substitution patterns confirm the possible influence of the f0 in S2 on the f0 in S1 (anticipatory coarticulation) by children suggested in the preceding analysis showing more errors in S1 than in S2. Three out of 5 of the remaining major substitution patterns for S1 involved erroneous production of T1, suggesting that T1 in S1 was more challenging for children. Six of the 17 major substitution patterns involved an incorrect tone in S2. Three of them involved substituting T2 with T3 (Appendix R). This finding was in accord with the findings in Wong et al. (2005), which reported T2/T3 confusions of children's monosyllabic productions by adult listeners and found more cases of T2 → T3 than T3 → T2 substitutions (Wong et al., 2005). T2 → T3 may indicate target undershoot of T2 in S2. Children might not have fully produced the rising portion of T2 in S2 by the end of the syllable, which could give rise to a percept of T3.

In summary children not only made more errors in their DT productions, their errors appeared to be less easily categorized by adults indicated by more substitution

patterns being used for children's productions. The major substitution patterns suggested that children's production of the tone in S1 was influenced by the tone in S2; children tended to modify the tone in S1 such that the offset of the f0 in S1 was close to the onset f0 for the tone in S2. T2/T3 confusions were the predominant errors in S2. Clearly, children's accuracy on tone productions was dependent on context; this issue was further examined in the next section on children's accuracy on the four Mandarin tones.

## Order of Accuracy Rates of the Four Mandarin Tones

No study has reported children's tone accuracy in different contexts, although children's tone data were usually collected in various contexts including monosyllabic to multisyllabic productions in various utterance positions. Children's tone accuracy rates were always reported for the four Mandarin tones with all the contexts collapsed. Thus, in order to compare the present results with those previous findings, we examined the order of the accuracy rates of the four Mandarin tones by (1) summing the accuracy rates across S1 and S2, (2) comparing the accuracy rates of the tones in each syllable position, and (3) comparing the accuracy rates of the tones in C and NC contexts in S1 and in C and NC contexts in S2. Finally, the accuracy rates of the same tone in different contexts were investigated.

As a first step, children's accuracy rates of the four tones were calculated collapsing across S1 and S2 and all contexts. Children as a group produced the four tones (T1, T2, T3, T4) with an overall accuracy rate of 82%, 76%, 79% and 78%, respectively. Figure 9A shows the boxplots of the scores. Freidman's ANOVA revealed no significant differences in the accuracy rates of the four tones when S1 and S2 were collapsed.

In light of the syllable effects on children's tone accuracy (see analysis above), the next step was to compare children's accuracy rates of the four tones in each syllable position. Figure 9B shows the boxplots of the children's accuracy rates in S1 and S2. Children produced the four tones in S1 with an accuracy rate of 75%, 80%, 75% and 83%, respectively; with no significant differences among the accuracy rates of the four tones. For S2, children as a group produced the four tones with 91%, 73%, 80% and 81%, respectively. Significant differences were found among the tones in S2 ($\chi^2(3) = 12.580$, N = 42, p = .006). Results of post hoc comparisons using Wilcoxon Signed Ranks tests revealed that children produced T1 better than T2 [z(N = 44) = -3.439, p = .001] and T3 [z(N = 42) = -2.922, p = .003]; T4 was significantly better than T2 [z(N = 44) = -2.002, p = .045] in S2, suggesting that the order of accuracy rates of the four tones in S2 from the highest to the lowest was: T1 ≈ T4 > T3 ≈ T2.

To investigate a possible interaction of tone compatibility and syllable position, the accuracy rates of the four tones in S1 under C contexts, in S1 under NC contexts, in S2 under C contexts and in S2 under NC contexts were analyzed independently. Figure 9C presents the boxplots of the distributions of the accuracy rates. Friedman's ANOVA showed nonsignificant differences across the four tones in all the comparisons except in S2 under NC conditions ($\chi^2 (3) = 10.936$, N = 35, p = .012) shown in the bottom right plot. Post hoc pairwise tests using Wilcoxon Signed Ranks tests showed that T2 was significantly worse than T1 [z (N = 41) = -3.237, p = .001] and marginally worse than T4 [z (N = 43) = -2.610, p = .009], after the p-values were corrected for multiple comparisons. The results suggested that T2 was the most difficult for children to produce in NC conditions in S2, and the order of accuracy rates of the four tones in NC contexts

in S2 in descending order was T1 ≈ T4 > T2, with T3 not significantly different than any of the other tones. Not all children provided data for all conditions. Thus, no age group comparisons were performed due to the lack of power.

Our next step was to compare the accuracy rates of the same tone in different contexts. When the accuracy rates of the same tone in S1 versus in S2 were compared, results of Wilcoxon Signed Ranks tests showed that T1 in S1 was significantly more difficult for children to produce than in S2, z (N = 42), = -3.523, p < .001. No significant differences were found in S1 vs. S2 for the other three tones.

More context effects were found when both the syllable position and the compatibility of tones were taken into account. Results of Wilcoxon Signed Ranks Test indicated that T1 was mostly influenced by syllable position. It was produced significantly better in S2 than in S1 in both C and NC contexts. T1 in S2 in C contexts was significantly better than T1 in S1 in C [z (N = 39) = -2.184, p = .029] and T1 in S1 in NC contexts [z (N = 40) = -2.801, p = .005]. The same was true for T1 in S2 in NC contexts. It was produced significantly better than T1 in S1 in C or NC contexts, z (N = 37) = -3.293, p = .001; and z (N = 39) = -2.813, p = .005, respectively. No significant differences were found for T1 in C versus NC contexts in either S1 or S2. Overall, the accuracy of T1 productions from highest to lowest by different contexts appeared to be T1 in S2 in NC contexts ≈ T1 in S2 in C contexts > T1 in S1 in C and NC contexts (Figure 9C).

T2 accuracy was more influenced by the compatibility of tones. In both S1 and S2, the accuracy of T2 was significantly better in C than in NC contexts, z (N = 37) = -2.236, p = .025 and z (N = 38) = -2.317, p = .021 for T2 in C vs. NC contexts in S1 and in S2,

respectively. T2 in C contexts in S1 was also significantly better than T2 in NC contexts in S2, z (N = 38) = -2.804, p = .005. No significant difference was found in S1 vs. S2 in either C or NC contexts. The order of the accuracy rates of T2 in different syllable and compatibility contexts seemed to be T2 in C contexts in S1 ≈ T2 in C contexts in S2 > T2 in NC contexts in S1 ≈ T2 in NC contexts in S2, although only the relations presented above were significant (Figure 9C).

For T3 and T4, the patterns of the accuracy in different contexts were similar and reflected the findings presented in the section on error patterns above. When children produced T3 and T4 in S1, they made significantly more errors when the tones were in NC than in C contexts, z (N = 38) = -2.878, p = .004 for T3; z (N = 42) = -2.647, p = .008 for T4. In NC contexts, the accuracy rates for T3 and T4 were significantly worse in S1 than in S2, z (N = 38) = -2.052, p = .040 for T3; z (N = 42) = -2.734, p = .006 for T4. The order of accuracy rates of T3 and T4 seemed to be: S1 in C contexts ≈ S2 in NC contexts ≈ S2 in C contexts > S1 in NC contexts (Figure 9C).

Taken together, the order of accuracy of the four tones was dependent on syllable position and tonal contexts. No differences among the four tones were found when S1 and S2 were collapsed and in S1 only. In S2, T2 and T3 were more difficult. When the tones were produced in NC contexts in S2, T2 was the most challenging for the children. T1 was most difficult in S1; T2 was most difficult in NC contexts. T3 and T4 were the most difficult in S1 in NC contexts.

## Correlational Analyses of Demographic Variables and Children's DT Production Accuracy

In this section Spearman's Rank Order correlations , $r_s$, were computed to determine whether children's performance in DT production was related to other variables of language performance and language experience. Then correlation analyses were performed to examine whether the number of target words the children produced was related to their DT accuracy rates. Lastly, the relation between the number of children who produced the DTs and the accuracy rates of the DTs was examined to explore possible effect of word familiarity on DT accuracy.

First, the relationship between children's Chinese and English educational background and DT accuracy rates was investigated. Nineteen children had attended Chinese preschools for one to 39 months (mean = 13.95, SD = 11.467) (Appendix A). A Spearman Rank-Order correlation coefficient was computed to examine whether duration of formal Chinese education was related to children's DT accuracy rates. The results showed that the overall accuracy rates on the DTs of these 19 children significantly correlated with the number of months they had attended Chinese schools, $r_s$ (N = 19) = .503, p = .028. $R^2$ = .25 (Appendix S)[2]. Twelve children attended English schools from one to 17 months (mean = 2.52, SD = 4.897). The number of months in English schools did not predict DT accuracy scores (Appendix S). Appendix T shows the scatterplots of the correlations of Chinese and English school education and DT accuracy rates.

Correlation analyses were then performed to investigate the relation of language scores and DT accuracy rates. No significant relation was found (Appendix S). Children's

[2] Number of months in Chinese schools was also found to be significantly correlated with age of the children ($r_s$ (N = 19) = .001, p = .001. $R^2$ = .44). Thus, the findings of significant correlation between number of months in Chinese schools and DT accuracy could have been confounded by age.

accuracy rates were not significantly correlated with their Chinese receptive percentile scores, Chinese expressive percentile scores, Chinese total percentile scores, English Receptive percentile Scores, English expressive percentile scores, or English total percentile scores.

Next, the relation between experience in Mandarin-speaking countries and DT accuracy was explored. Twenty-seven children had visited and stayed in their native countries (25 in China, 2 in Taiwan) for one to 78 months (mean = 22.15, SD = 19.06). No significant relation was found between the number of months in the native country and DT accuracy rates (Appendices A and S).

There was also no significant relation between the number of words produced and DT accuracy rates. Children who produced more words (words that were in the target form but were excluded due to noise or in non-isolated positions were included for this analysis) did not attain higher production accuracy rates (Appendix S).

Finally, a Spearman's Rank Order correlation was used to investigate whether words that were produced by more children were produced more accurately. All the 30 High and Low words were included for investigation. Frequency of productions was defined as the percent of children in this study and in the pilot study on word familiarity who produced the target words. No significant relation was found, $r_s$ (N = 30) = .109, p = .568. The relation remained non-significant when children in the pilot study were excluded, $r_s$ (N = 30) = .233, p = .216.

Overall, despite the differences in some of the demographic variables of the children and different number of target words the children produced, none of the factors examined above had significant effects on children's DT production accuracy except that

the longer the children had been in Chinese preschools, the better their DT production. Words produced by more children did not predict DT accuracy.

# Chapter 3: Acoustic Analysis of Tones

To gain more information about the f0 contours produced by children when attempting DT productions and make inferences about children's articulatory movements during the production the DTs, acoustic analysis was performed on selected adults' and children's productions.

## Method

### Stimuli

Three sets of stimuli were chosen for acoustic analysis. The first two sets consisted of adults' and children's correct DT productions of High Words, in which the tones were correctly identified by all three judges. The third set of stimuli involved children's incorrect DT productions in which the productions were unanimously misidentified as the same (substituted) DT pattern by all three judges.

#### Correct Adult Productions

The productions of the DTs in High Words by three adults (UA22, UA53, and UA68) were selected. All three speakers produced all 15 High Words, and all their productions were usable and included in the study. UA22 and UA53 were selected because all their High Word productions were correctly identified by the three judges. UA68 was randomly picked from the adults whose DT productions elicited only one judgment error (i.e., 44 correct out of 45 judgments on the 15 DTs by the 3 judges). T11

produced by UA68, which was misidentified as T41 by one judge, was excluded. Altogether 44 (15 x 2 + 14) productions by the three adults were included.

Correct Child Productions

For each of the 15 High Words, two correct child productions with the tones correctly identified by all three judges were selected. One of them was randomly chosen from the correct productions by two-year-old children and the other was randomly chosen from 5- and 6-year-olds. For T12 and T14 no correct productions by 2-year-olds were found. Thus, a correct production produced by three-year-olds was randomly selected for these two DTs. Altogether 30 correct child productions were included.

Major Substitution Patterns of Children's Productions

Children's productions that represented the major substitution patterns (i.e., substitution patterns that accounted for more than 10% of the total judgments for the target DT) that the judges used to categorize the incorrect DTs produced by C2 to C4 children (highlighted in Table 6 and listed in Appendix R) were selected. For each major substitution pattern, children's High Word productions that elicited the same substitution error by all three judges were selected. For example, for T11→ T21, all three judges identified the tones of the High Word <xi1gua1> produced by UC01 and UC35 as T21. Thus, these two productions were selected. No child production elicited a consistent substitution error by the judges for four of the 17 major substitution patterns (T31→ T21, T31→ T41, T34 → T24 and T42 → T43). Most of them (3 out of 4) involved T3 in S1. Only one child production was found for T11→ T41, T14 → T24, T14 → T44, T24→T23, and T43 → T13. For the other eight major substitution patterns, two to seven

child productions were included. Altogether 32 productions that represented 13 of the major substitution patterns were analyzed.

## Acoustic Analysis

### Segmentation

Because acoustic-phonetic information about the consonants was largely eliminated in the filtered stimuli, segmentation of the productions was performed on the original (unfiltered) stimuli using a custom written script (Xu, 2008) for PRAAT. (Boersma & Weenink, 1992). The waveform, the spectrogram, and a label window were shown on the screen. The onset and offset of the consonant and rime for each syllable were marked and labeled manually. Thus, the DT was divided into four segments (C1, R1, C2, R2), for which f0 values and duration information were obtained. C1 started from the beginning of the initial consonant to the onset of the first vocal pulse for the vowel in S1. The beginning of the initial consonants for the stops, fricatives and affricates, nasals, and approximants was defined as the onset of the release burst, the fricative noise, the nasal murmur and the formant transition, respectively. R1 started at the end of C1 (i.e., the onset of the first vocal pulse for the vowel in S1) and terminated at the cessation of the last vocal pulse of the vowel in S1, that corresponded to a sharp decrease in the amplitude in the waveform for the vowel in S1. In syllables with a final nasal, R1 ended at the end of the nasal murmur. C2 started at the end of R1 and ended at the onset of the first vocal pulse for the vowel in S2. In cases where the initial consonant for S2 was a nasal or glide, the segment for C2 included the nasal murmur for the nasal and the formant transition for the glide, respectively. For the word <dian4nao4> where the final consonant of S1 and the initial consonant of S2 were both the nasal /n/, the end of R1 and the beginning of C2

was defined as the midpoint between the end of the vowel for S1 and the beginning of the vowel for S2. R2 started at the onset of the first vocal pulse for the vowel in S2 and ended at the cessation of the last vocal pulses for the vowel in S2, corresponding to the end of the waveform or a sharp decrease in the amplitude of the waveform for the vowel. In cases where R2 ended with a nasal, the end of R2 was the end of the last nasal bar for the nasal murmur.

Duration

The durations of the four marked segments (C1, R1, C2, R2) were measured and saved in a text file. Each segment interval was divided into 10 equal time intervals. Thus there were 40 time segments (equal in interval within each of the 4 segments—C1, R1, C2, R2) for each DT.

Vocal Pulses

F0 extraction and measurements were done on the filtered tokens using the same custom written PRAAT script (Xu, 2008). In the label window, the onset and offset of the consonants and rimes were marked based on the information obtained in the unfiltered counterpart (see the previous paragraph). Vocal pulse markings generated by PRAAT were inspected for any erroneous markings (missing pulses and double markings) and corrected manually. The script divided each of the four segments (C1, R1, C2, R2) into 10 intervals equal in time, computed the mean f0 for each interval, applied a smoothing algorithm to remove local spikes in the f0 values, and saved the f0 values in a text file for each word by each speaker. Thus, for each DT, there were 10 f0 values for each of the four normalized segment intervals in each DT.

## F0 Plots

The f0 trajectories based on the measurements of the 40 intervals in adults' and children's productions were plotted for comparison. To facilitate comparisons of the shape of the f0 contours across productions and speakers, the 40 f0 values for each DT production were plotted with equal intervals on the abscissa (See Figures 10–11). Thus, when interpreting the f0 contours, one thing to note is that although the intervals for the consonants and rimes were presented with equal time interval value, the consonants were actually much shorter than the rime. Because no vocal fold vibrations were involved in voiceless consonants, no f0 values or f0 tracks were presented in all C1 and C2 except for the ones with nasals and approximants. Some DTs were produced with creaky voice (glottal fry) by the children and adults. Glottal fry occurred when the speaker produced the tones with extremely low f0 and caused the vocal folds to vibrate slowly with extreme jitter (unequal vocal periods). Due to measurement issues with glottal fry, it is not easy to compare f0 contours for productions with and without glottal fry. Thus, comparisons of f0 in the following sections will focus on f0 contours that did not contain any glottal fry, although some f0 contours with glottal fry arre presented in Figure 10.

## Results

### Correct Adult Productions

The left panels in Figure 10 show the f0 contours of the correct productions by the three adults with time normalized (f0 values were obtained and plotted at equal intervals within C1, R1, C2, and R2). The f0 contours of many of the DTs produced by the three adults were similar in terms of f0 height, f0 change, direction and slope of the f0 contours.

The f0 contours that seemed to have some speaker variations were due to glottal fry. All the T3 in S2 by all 3 adult speakers, most of T4 in S2 (T14, T24, T44) by all 3 speakers and T2 in T12 by UA53 were all produced with different degrees of glottal fry. Overall, adults produced the DTs with similar f0 contours.

Correct Child Productions

The right panels in Figure 10 show the f0 contours of the correct productions by the selected children with time normalized (marked with triangles in the contours). The f0 contours for the same DTs by an adult (UA22) were also plotted in the charts (marked with diamonds in the contours) for easy comparisons between the f0 shapes of the adult's and children's DTs. UA22 was selected because she produced all the 15 DTs, all her productions were correctly identified by all three judges and the f0 contours of her DTs were the least affected by glottal fry. The codes in the legend in the charts provide information on the child who produced the DT. For example, "UC22_C5Corr" represents a production by the child with the ID number "UC22". C5 indicated that the child was in the C5+ group. "Corr" represented that it was a correct production of the DT (all 3 judges identified the target DTs). Like adults, children produced the DTs with creaky voice in T2, T3 and T4 in S2 (T12, T13, T23, T43, T14, T24, T44), although mostly by one of the two children only. However, unlike the adults, children also produced T3 in S1 with glottal fry (one incident in T31 and T34).

When compared to adults' f0 contours (left and right panels in Figure 10), some variations in the f0 contours of children's productions were observed. Children's f0 contours of T1 in S1 in NC contexts (T12 and T13) were not consistently produced with a steady high and flat f0 contour like in the adults' (Times 10-20 in Charts B2 and B3 in

Figure 10). T2 was sometimes produced with a reduced upward slope in S1 (T21, T22, T24) and S2 [T22, T32 (by one child only), T42] and a higher f0 at the onset of S2. The f0 differences at the syllable boundary of T22 were also reduced. These results suggested that children tended to produce T2 in these contexts with reduced f0 changes. T3 in S1 remained low in f0 in adults' productions; however, the f0 contours of T3 in S1 in NC contexts (T31 and T34) in children's productions started to rise at the end of S1, indicating anticipation of the tone starting with a high f0 in S2. T3 in S2 in T23 produced by UC52 did not go as low as the adults' f0 contours. Compared to the adults, children produced the f0 contours with a reduced downward slope in T4 in S1 (T41, T42, T43, T44) and in S2 (T24 and T34 by the C5 child). Overall, although children's productions were correctly identified by all three judges, some of the f0 contours were different from those of the adults. Anticipatory coarticulation, simplification of f0 complexity, and target undershoot were observed.

Selected Substitution Patterns

Our next step was to examine the f0 contours in children's incorrect productions. To facilitate comparison of the f0 contours, nine of the 32 child productions (T22 → T23 by UC29, UC55 and UC48, T23 → T13 by UC39, UC67 and UC70, T41 → T11 by UC33, T43 → T13 by UC36, T43 → T42 by UC42) that represented the major substitution patterns were excluded due to the presence of glottal fry (creaky voice), which had very different f0 contours that made comparisons very difficult. To reduce the number of f0 contours in each chart, one production of T42 → T41 and three productions of T44 → T14 were also excluded such that the number of f0 contours for incorrect child productions in each chart would not exceed two. The f0 contours of these four

productions were qualitatively similar to the ones included for the same substitution patterns. Altogether, the f0 contours of 19 children's productions representing 12 substitution patterns were presented and compared.

Figure 11 presents the f0 contours of the 19 child productions. The correct adult target forms of the substituted DTs (e.g., T11 in T11 → T21) produced by UA22 (marked with diamonds in the darkest line) and the perceived DT (e.g., T21 in T11 → T21) produced by UA22 (marked by the lightest line) were included to determine how children's f0 contours were different from the target form and how children's f0 contours resembled the f0 contours of the perceived DTs in the adultforms. The numbers in the parentheses in the legend code indicate the age of the child. For example, 208 represents 2 years and 8 months old.

Seven out of 12 of the children's substitution patterns presented in Figure 11 involved simplification of the f0 contours in the target DTs. Several of them involved the simplification of the f0 contours by moving the f0 contours in S1 towards the f0 onset of the tone in S2 (anticipatory effect). In children's T44 → T14 (Figures 11A and 11B) and T41 → T11 (Figure 11C) productions, the f0 contour in T4 in S1 was reduced to essentially a flat contour with the f0 level close to the onset f0 of T4 in S2. In T11 → T41 production (Figure 11D), T1 in S1 went slightly downward towards the f0 onset of T1 in S2. In T14 → T24 (Figure 11E), T1 in S1 went slightly upward toward the initial f0 in S2.

Simplification of f0 occurred in S2 in T12 → T13 (Figure 11F). For this DT, children's f0 contours in S2 had no sharp f0 turns as in the adults'. The f0 slope and f0 range in S2 were much reduced.

Some children's substitution patterns involved simplification of f0 in both syllables. In T42 → T41 (Figure 11G), the slope of T4 in S1 and T2 in S2 were both flattened. In T23 → T13 (Figure 11H), the f0 contour of T2 in S1 had a much reduced rising slope and the rising component of T3 in S2 was much reduced. In T22 → T23 (Figure 11I), T2 in S1 was produced with a flattened f0, the slopes of f0 change (degree of downward and upward movements) in T3 in S2 were reduced, and the difference between the f0 offset of S1 and f0 onset of S2 was reduced.

In the case of T11 → T21 (Figure 11J), the relative height of the f0 contours of T1 in S1 versus S2 seemed to contribute to the misperception of children's T11 as T21. As shown in Figure 11J, the shape and height of the children's f0 contours of T1 in S1 were more similar to the adults' T1 in S1 of T11 than T2 of S1 in T21. However, adults produced T1 in S2 with a slightly lower f0 than that in S1 (compare figure 10 A1 and figure 11J), whereas children produced T1 in S2 with a higher onset f0.

No obvious f0 simplification was found in two other substitution patterns (T14 → T44, T24 → T23). In T14 → T44 (Figure 11K) and T24 → T23 (Figure 11L), the children's f0 contours were similar to the f0 contour of the misperceived DTs (i.e., T44 and T23) produced by adults, suggesting that these substitution patterns were more likely due to different perceptual representations of the tones for the target words. For the last substitution pattern, T43 → T42 (Figure 11M), the child's production might be characterized as target undershoot of the tone in S2 (i.e., it did not go as low as it should be, another form of f0 simplification) or had perceived the tones of the target word as T42 because the f0 contour of the child's production resembled adults' T42.

Overall, most of the substitution patterns presented here showed simplification of the f0 contours of the target DTs by reduction in the speed of f0 change (flattened f0 slope), reduction in the f0 difference at the syllable boundary of S1 and S2, and/or reduction in the change of f0 direction (decrease in the number of f0 turns). A couple of substitution patterns suggested possible misperception of the DTs in the target words by the children.

# Chapter 4: Discussion

As stated in the introduction, the two main goals of the present study were to answer the questions: 1) how do Mandarin children's productions of lexical tones in familiar disyllabic words develop over time and 2) are disyllabic tones with more complex f0 contours (non-compatible tone combinations) more difficult to master? In this section the findings related to these questions are summarized and discussed with respect to previous research on the development of lexical tone production. In addition, the effects of syllabic context and complexity of tone contours on children's tone production accuracy and correlations between children's demographic data and disyllabic tone production accuracy are presented.

The following discussion was based primarily on the findings for the analysis of the 15 High Words (one lexical item for each disyllable tone combination) for several reasons. First, children's developmental trends in overall production accuracy based on the analysis of all 30 words (two words for each tone combination) and the High Words were very similar. Second, a large proportion (34%-39%) of the children in the younger age groups produced only one of the two words for each target tone combination (Table 2) and one-third of the Low Words were produced by only three to four children. These lexical effects led to different numbers of judgments for young versus older children being combined in the analysis of developmental trends in the production of particular tone combinations. To avoid this possible confound, the High Words analysis was used to examine the role of tone contour complexity and syllable context effects on production mastery.

## Developmental Trends of Children's Disyllabic Tone Accuracy

Most previous studies on children's acquisition of lexical tones reported that lexical tones were acquired very early, before segmental production was mastered (Clumeck, 1980; Clumeck, 1980; Li & Thompson, 1977). For instance, recent studies in which tone judgments of children's productions were based on natural (unfiltered) stimuli reported that tone productions were stabilized before two years of age (Hua, 2002) and that children as young as one and a half years old produced no tone errors (Hua & Dodd, 2000) in monosyllabic to trisyllabic words in isolation and in spontaneous and elicited continuous speech. However, a recent study which determined children's tone accuracy using filtered speech to eliminate lexical information available to the native Mandarin listeners who judged tone accuracy reported that three-year-old children did not produce the four Mandarin tones in isolated monosyllabic words as accurately as adults (Wong et al., 2005). Results of the present investigation support the latter finding and indicate that lexical tones are not mastered as early as most studies have suggested; in general, most children do not master tone production in disyllabic words before six years of age.

To answer our first question—how do children develop their disyllabic tones over time, findings in terms of the age related changes in the overall accuracy rates of children's disyllabic tones, the accuracy rates of the 15 disyllabic tones, variability in children's accuracy rates, and number of judges who miscategorized children's disyllabic tone productions in the High Words are summarized, and possible reasons for these developmental trends are proposed. Regression in overall accuracy in the performance of four-year-olds and lexical effects on children's tone acquisition are also discussed.

Development of Disyllabic Tone Production

As a group, the two- to six-year-old children tested in this study did not produce the disyllabic tones with adult-like accuracy. Their overall accuracy rates, collapsing over the 15 disyllabic tones, were significantly lower than for adult productions, filtered and judged in the same way by the same native Mandarin listeners. Children's disyllabic tone productions improved significantly with age (Figure 6B); age accounted for approximately 21% of the variance in the growth of 2- to 6-year-old children's disyllabic tone accuracy. At two years of age, individual children's productions of tone combinations were judged as correct between 28% and 88% of the time (mean accuracy = 52%). In three year olds, overall accuracy scores ranged from 40% to 90% (mean accuracy = 67%). The overall accuracy scores for four-year-old children tended to be lower than those for three-year-olds, with their overall accuracy rates between 30% and 80% (mean accuracy = 60%). Children five years or older showed significant improvement in overall accuracy, ranging from 70% to 100% (mean = 85%); on average this group was significantly better than two- and four-year-old groups. Two 5-year-old children produced the disyllable tone combinations with accuracy rates approaching adults' performance. One other five-year-old child in this study attained an overall accuracy rate within the 95% confidence interval of the adults' scores. This child's performance may not be typical for her age because she was reported to be very advanced in her speech and language skills by her parents and teachers.

Two- to six-year-old children as a group produced 13 of the 15 disyllabic tones (except for T11 and T14) with accuracy rates significantly lower than adults' (Figure 4B). Accuracy rates on individual disyllabic tone combinations generally improved across age

groups, although the rates of growth varied for different tone combinations. Disyllabic tone contours that were produced with relatively high accuracy by two-year-old children (T24, T13) were not the best mastered DTs by the five- to six-year-old children (Table 4). Overall, T11, T21 and T43 underwent the most developmental growth, whereas T24, T31 and T32 showed the least improvement across the age groups. By the age of five, children produced eleven of the fifteen disyllabic tones (all but T12, T13, T24, and T44) with over 80% accuracy.

If age of mastery of disyllabic tone production is defined by the number of children that produced the disyllabic tones with adult-like accuracy, T21 and T32 were mastered by the most two- to six-year-old children (bottom half of Table 3); T12 was mastered by the fewest children, followed by T22 and T44. DTs that were mastered by the most children in the four age groups had compatible tone combinations. On the other hand, most disyllabic tones that were mastered by the fewest children in the four age groups had non-compatible tone combinations. The 2-year-olds had difficulties with both compatible and non-compatible tone combinations. The effects of tone complexity on production accuracy are discussed further below.

There was some evidence of regression in accuracy rates by four-year-old children. Fewer four-year-olds produced the disyllabic tones with adult-like accuracy than three-year-olds. Nine of the 15 disyllabic tone combinations were produced with adult accuracy by more three-year-olds than four-year-olds.

As shown, the mastery of disyllabic tone production is a lengthy and gradual process; even 5- to 6-year-old children's productions have not met adult standards. This protracted course of development can be explained by the general processes of speech

production development. The acquisition of lexical tones, like segmental speech sounds, is a complex process that involves an integration of development in neurological, physiological, and cognitive systems (Kent & Vorperian, 2007; Locke, 1983), and relies on the anatomical maturation and physiological proficiencies of various biological systems (Kent, 2000; Kent, 2004; Walsh & Smith, 2002).

For accurate production of tones, the child has to, first, establish accurate phonological representations of the tones. It takes time for the child to be exposed to sufficient speech input produced by different speakers in various contexts so that s/he will be able to extract relevant phonetic information from the acoustic signal. Little research has been conducted on children's tone perception, and most of this research has investigated the perception of tone contrasts in isolated, monosyllabic utterances, probably due to the fact that very few minimal pairs are found in young children's disyllabic and multisyllabic vocabulary. Wong et al. (2005) reported that 3-year-old children attained high accuracy rates in identifying the four Mandarin tones in familiar monosyllabic words (Wong et al., 2005). In the trials that involved minimal pairs in the study, children perceived T1, T2, and T4 with accuracy rates higher than 80%. T3 was identified with slightly lower accuracy (69%). Given that the study only examined perception of monosyllabic tones, children's ability to distinguish the tones in different contexts or at different ages remains unclear.

Even if young children have acquired good perceptual skills for lexical tones early on, they still have to learn to control and coordinate complex articulatory movements that regulate the laryngeal and supralaryngeal structures to produce the tones and the segmental speech sounds precisely, efficiently and simultaneously. Such a perception-

action linkage takes time to develop and requires the support of a mature neuro-musuclar system (Kent & Vorperian, 2007). However, the central neural mechanisms for motor timing control and the neuromuscular capabilities for speech motor control do not stabilize or reach maturity until after 10 years of age (Kent & Vorperian, 2007; Tingley & Allen, 1975).

To complicate the situation, the anatomical structures and physiological proficiencies of children are still developing in young children (Kent & Vorperian, 1995; Ostry, Feltham, & Munhall, 1984); consequently, young children have to learn to produce speech sounds while their speech production systems are undergoing substantial anatomical and physiological changes. The size and shape of the vocal tract change considerably in the first few years of life (Kent & Vorperian, 1995). The laryngeal structures, which have a direct impact on children's tone production, also undergo rapid changes. The larynx increases in size and descends during development. At birth, the larynx is about 4-5 mm long and positioned at about the 2$^{nd}$ cervical vertebra. By age 5 years, larynges grow to approximately 7.5 mm and are at the level of the 4$^{th}$ and 5$^{th}$ cervical vertebra (Kent & Vorperian, 1995; Vorperian & Kent, 2007). The morphology and composition of the vocal folds also change appreciatively (Hirano et al., 1983). In newborns, the vocal folds have a uniform structure without the three distinctive layers found in adults'. Between the age of one and four years, the intermediate layer, which consists primarily of elastic fibers, and the deep layer, which is composed of collageous fibers, of the vocal folds evolve. With time, the amount of microfibrils in the vocal folds decreases and the number of amorphous components increases, which result in more elastic vocal folds (see review in Kent & Vorperian, 1995 & Kent & Vorperian, 2007).

These developmental changes in the larynx would cause different vibration patterns and acoustic output of the vocal folds. Thus, during development, the child needs to constantly learn to use different articulatory gestures to produce the same speech targets (Callan, Kent, Guenther, & Vorperian, 2000).

In all, accuracy in production of lexical tone sequences requires phonetic learning, motor learning, and implementation of the motor plan with precise, efficient and temporally coordinated laryngeal and supralaryngeal articulatory gestures. It relies on the maturation of multiple anatomic, neuralgic and physiologic systems (Smith & Zelaznik, 2004). Given that these systems are still developing in the young child, children's acquisition of tones may be constrained by their capabilities to control these developing structures. Thus, it is not surprising that adult-like tone production takes time to master.

Age Related Changes in Inter-subject Variability

Substantial across-subject variability was observed in the 44 children's disyllabic tone productions in terms of overall accuracy rates summing over the 15 disyllabic tones (Figure 6B) and in individual disyllabic tones (Figure 4B). When comparing across the age groups, some developmental changes were observed in children's variability in the accuracy rates. Inter-subject variability in overall accuracy, collapsing over all the disyllabic tones (Figures 3B & 6B), and in most of the 15 disyllabic tones, declined across age groups (Figure 5B), with the inter-subject variability decreasing most substantially in five to six-year-old children (Figures 5B & 6B).

Developmental variability in speech sound production has been widely reported in the literature and is interpreted as an indication of neuromotor immaturity (Kent, 1976; Stathopoulos, 1995). Young children's speech motor performance has been found to be

less consistent, more unstable, and more variable than adults (Green, Moore, Higashikawa, & Steeve, 2000; Kent, 1992; Ostry et al., 1984; Smith & Zelaznik, 2004). They produce articulatory gestures that are less coordinated, less precise in timing, more limited in rate, with more restricted range of movement, and with longer movement durations (Green et al., 2000; Kent, 1992; Walsh & Smith, 2002), resulting in more variable segment, syllable, and phrase durations (Smith, 1991), vowel formant frequencies (Eguchi & Hirsch, 1969), lip rounding (Goffman, Smith, Heisler, & Ho, 2008), neural commands to muscles (Wohlert & Smith, 2002), articulatory movements (Stathopoulos, 1995) and coordination of articulatory gestures (Wohlert & Smith, 2002). As children get older, they progressively refine their speech motor control and the coordination, timing and placement of their articulatory gestures become more precise (Kent, 1976). Thus, inter-subject and intra-subject variability decrease with age.

Variability of speech motor movements may also be related to anatomical immaturity. Lecours (1975) posited that the immaturity in learning the motor patterns of speech production by young children was related to incomplete myelogenesis of different structures and pathways in the central nervous system. For example, the myelination of the fasciculus arcuatus in the associate bundles in the cortices, which is responsible for transmitting auditory information to the motor areas for speech production, was not complete by the age of two years; the myelogenesis of the axial fibers of the angular and supramarginal gyri in the associate bundles also does not reach maturity before the age of six or seven years.

Another theoretical perspective on children's variability in speech production claims that speech variability is a result of an adaptive mechanism in the developing child

(Wohlert & Smith, 2002). According to this view, the young child is required to remain flexible in his/her articulatory movements, and constantly explores and learns different articulatory patterns to achieve the same articulatory goal, in response to the anatomical and biomechanical changes in his/her speech production system (Callan et al., 2000; Stathopoulos, 1995).

All of the above factors that contribute to segmental variability would also be expected to affect the laryngeal movements associated with tone production. Thus, similar to the acquisition of segmental speech sounds, substantial inter-subject variability was observed here in the course of the development of disyllabic tone production. The variability diminished appreciably by the age of five years. Given the design of the present study, no information on intra-subject variability on tone production was provided. Future studies on children's repetition of disyllabic tones could provide a more complete account of children's variability in disyllabic tone production.

Age Related Changes in the Number of Listeners Who Mis-identified Children's Disyllabic Tones

Younger children's tone productions tended to be mis-identified by more judges (Table 5) and the judges tended to use more response alternatives to categorize younger children's disyllabic tones than older children's (Appendices N-Q). Similar to the patterns found in accuracy rates and inter-subject variability presented above, five-year-old children's disyllabic productions yielded a substantial decrease in the number of judges who mis-identified the tones and the number of different response alternatives the judges used to categorize a particular production. These developmental trends may be indicative of more precise disyllabic tone production by older children than younger

children. If younger children produced f0 contours that deviated significantly from any of the canonical target f0 contours for the four tones, the judges would have more difficulty categorizing the tones and would have less consistency in selecting the same responses for each DT production. More extensive acoustic analysis on children's incorrect disyllabic tones that were categorized as different disyllabic tones by different judges will provide more information on young children's disyllabic tone patterns and how children approach the adult forms over time.

Apparent Regression in Four-year-old Children

The data reported here suggest some regression in four-year-old children's performance. Four-year-old children produced the disyllabic tones with a lower mean overall accuracy than 3-year-olds; and the best performers in the 4-year-old group had lower overall accuracy scores than the best performers in two- and three-year-old groups (Figure 6B). Nine of the 15 disyllabic tones were produced with adult-like accuracy by fewer 4-year-old than 3-year-old children (Table 3). More 4-year-old than 3-year-old children's productions were misidentified by two or more judges (Table 5), and more substitution patterns were used for categorizing 4-year-old children's disyllabic tone productions than for 3-year-olds' (Appendices O & P).

Several reasons can be hypothesized for the regression in the performance of 4-year-old children. First, it may be a consequence of anatomic changes in 4-year-old children. As stated above, children undergo considerable changes in the laryngeal structures, particularly the vocal folds, between 1-4 years of age. Such radical physical changes in the larynx might interrupt children's development in disyllabic tone production.

Another possible reason for the decline in performance accuracy by 4-year-old children might be related to attention. Children develop different speech production skills at different rates and different times. Challenges, growth spurts or changes in other areas or domains of development (e.g., syntactic development, motor development, second language learning, formal education) might draw the child's attention away from tone production and cause a temporary decline in performance. Further analysis of the productions gathered here in terms of the correlations between segmental production accuracy and tone production accuracy will be done to address this issue in more detail.

The reported findings might be due to the small sample size. Only 11 four-year-old children were included, so the results might reflect sampling errors. However, in a previous (as yet unpublished) study, a similar regression in disyllabic tone accuracy in 5-year-old Mandarin-speaking children growing up in Taiwan was observed. In any case, future larger scale studies and longitudinal studies are needed to confirm the presence of a regression in disyllabic tone accuracy in 4 to 5-year-old children and the relationship between the development of tone and other aspects of phonological developments.

Lexical Effects on Children's Disyllabic Tone Development

The data reported here suggest that children's productions of tone contours in familiar disyllabic words differed as a function of the specific lexical items that they were attempting. In some cases, they produced the same tone combination in one word better than in the other. Many of the lexical effects exhibited by 2-year-old children could be attributed to small sample sizes in attempted utterances, and having mostly different speakers producing the two different lexical items with the same disyllabic tone contour. However, lexical effects in the High versus Low words for several disyllabic tones

exhibited by the children across the age groups were produced by the same five or more speakers (compare Figure 7A and Appendix K), and cannot be readily explained by the two factors suggested for the 2-year-olds. Future research needs to look systematically at children's productions of disyllabic tone contours across different words with different segmental content and different familiarity to shed some light on the relations between lexical familiarity, segmental complexity, and tone contour production accuracy.

Summary

In summary, disyllabic tone acquisition is a gradual process that spans over 6 years in Mandarin-learning children growing up in a relatively monolingual environment in New York City. Children as old as six years of age did not produce disyllabic tone combinations in an adult-like manner. Younger children seemed to produce disyllabic tones with f0 contours that deviated more from the adult forms. There was significant improvement in overall judged accuracy of disyllable tone contours from 2 to 6 years, with large intra-group variability among 2- to 4-year olds. Accuracy rates varied across the 15 disyllabic tone combinations. As children became older, their accuracy rates improved; however, different tone combinations progressed at different rates. Several disyllabic tones remained difficult even for five- and six-year-olds. Four-year-old children appeared to regress in overall tone production accuracy, whereas five- to six-year-old children demonstrated significant improvement in accuracy and smaller inter-subject variability. There were possible lexical effects on children's disyllabic tone development.

## F0 complexity and Children's Disyllabic Tone Production

At the outset of the study, we hypothesized that young children would have more difficulty with tone combinations that had more complex f0 contours. More complex f0 contours involve relatively rapid changes in velocity and direction of f0 that may be difficult for children to produce given the immaturity of articulatory gestural control. Specifically, we predicted that non-compatible tone combinations, which have large f0 differences between the offset of the tone in the first syllable and the onset of the tone in the second syllable, would be more difficult for children to produce than compatible tone combinations, which have smaller f0 differences at the boundary between the two syllables. Also, tones in the second syllable of disyllabic tones were predicted to be more difficult for young children. In adult disyllabic combinations, the tone contour in the first syllable varies less from its canonical form than the tone contour in the second syllable (Xu, 2001). That is, adult tone contours exhibit more carry-over coarticulation. Thus, the second syllable includes the transition of f0 from the first to the second tone and the tonal target of the second tone is often realized in the latter half of the vocalic rime. Implementing more complex f0 contours within the syllable time frame should be more challenging for more immature articulatory systems.

The findings in the present study supported our hypothesis that disyllabic tone combinations with more complex f0 contours were more difficult for children to produce. Children's overall accuracy rates for non-compatible tone combinations were significantly lower than for compatible tone combinations. Moreover, the non-compatible tone combinations were among the contours that were produced by the fewest number of 3- to 6-year-old children with adult-like accuracy (Table 3). In contrast the most difficult

disyllabic tones for two-year-old children included both compatible and non-compatible tone combinations. Together with the findings that 2-year-old children had low accuracy rates on most of the disyllabic tones, the results may indicate that 2-year-old children have difficulties producing many of the disyllabic tones even when the f0 contours are rather simple.

The unexpected finding that children's tone productions in the second syllable were more accurate than in the first syllable was not predicted, but does not refute our hypothesis about f0 complexity and accuracy rates. Detailed analyses showed that when children produced non-compatible tone combinations, they made significantly more errors on the tones in the first syllable than in the second syllable. This tendency was observed in all age groups although only 4-year-old children reached statistical significance due to lower power. The major substitution patterns of children's errors suggested that when children produced NC tone combinations, they tended to modify the f0 contours in S1 such that the f0 shift at the boundary between S1 and S2 was reduced (Appendix R). The f0 tracks in the acoustic analysis also supported such observation (Figure 11). These findings together suggest that children had more difficulties with non-compatible tone combinations, and their strategy in producing these contours was to modify the f0 contour of the tone in the first syllable to reduce the f0 difference at the syllable boundary between S1 and S2. One possible reason for this is that children may anticipate an f0 shift at the syllable boundary which would cause greater demands on the articulators. Therefore, they try to reduce the demands by starting movement towards the tonal target for the second tone earlier, in the first syllable, even when such gestures may

cause a different tonal percept of the target tone in the first syllable. This strategy reduces rapid f0 changes in the upcoming tone.

If children modify the f0 contours in S1 to reduce the f0 difference at the syllable boundary between S1 and S2 while producing NC tone combinations, it is not surprising to find that children produced tones closer to canonical targets (as judged by listeners) in S2 than in S1. Given that in NC contexts, the f0 contour in S1 was modified such that the f0 difference at the syllable boundary was reduced, the tone in S2 would become more compatible to the tone in S1. Therefore, there would be less difference between the f0 contours for the same tone in C versus NC contexts in S2. This explains our findings that children made comparable errors in C and NC contexts in S2 as well as in C contexts in S1.

Children did not only modify the f0 contours in NC tone combinations, results from the analyses of the major substitution patterns by children and preliminary acoustic analysis showed that children tended to simplify the f0 contours in both compatible and non-compatible disyllable tone combinations, although the tendency was more obvious in non-compatible contours. This was observed in both children's correct and incorrect productions. In children's disyllabic tone productions that were judged as the intended patterns (correct items), the f0 contours were produced with reduced f0 slope, smaller f0 differences at the syllable boundary, and undershoot of the tone target (e.g., the f0 of T3 did not go as low as it should) (Figure 10). Acoustic analysis of children's major substitution patterns showed greater degrees of f0 simplification in children's incorrect than correct disyllabic tone productions. Reduction of f0 slopes and undershoot of tonal targets were observed mostly in the first syllable, although they were also found in the

second syllable. In a number of cases, the f0 contours in the first syllable moved towards the f0 onset of the tone in the second syllable and exhibited an f0 contour that was very different from the (adult) canonical target f0 contour in the first syllable (Figure 11).

These findings consistently showed that children tended to simplify the f0 contours and had more difficulties producing more complex f0 contours. Two possible reasons can be provided for children's simplification of the f0 contours and higher error rates in disyllabic tone combinations with more complex f0 contours. First, given the rapid f0 change in complex f0 contours, children may have more difficulty perceiving the relevant phonetic information in the acoustic signal, and may, therefore, require more time and experience to establish accurate phonetic representations of more complex tone combinations, particularly for words with non-compatible tone combinations.

Another possible reason for children's greater difficulty in producing more complex disyllabic tones may be due to physiological and motoric constraints in the developing systems of the child. In adult productions, the realization of the tonal target is accomplished within the syllable. More complex f0 patterns entail more rapid and, temporally coordinated changes in the speed, acceleration and deceleration of the articulatory gestures. Given that children are reported to have slower and less precise speech motor control (Goffman & Smith, 1999; Walsh & Smith, 2002), there is reason to speculate that children have more difficulties producing more complex f0 contours due to reduced motoric and physiological capabilities.

The fact that children made comparable numbers of tone errors on compatible versus non-compatible disyllables in the second syllable and produced the tones more accurately in compatible versus non-compatible contexts in the first syllable suggests that

the bigger contributing factor to children's disyllabic tone errors lay in production immaturity rather than incorrect or underspecified perceptual representations. In adult's disyllabic tone productions, the f0 transition takes place in the second syllable. Therefore, the f0 contours in the second syllable are more complex and variable (further from canonical forms) than in the first syllable, especially when the tones in the first and second syllable are non-compatible. If children had more difficulty perceiving complex, coarticulated f0 contours and, consequently, had less accurate representations of tones with more complex f0 contours, they should have had more tone errors in the second syllable, particularly in non-compatible disyllable tone combinations. Moreover, in adult speech, the f0 contours of the same tone in the first syllable are very similar across different contexts, regardless of what the upcoming tone is (Figure 2). Thus, it is less likely that children would have more difficulties perceiving the tones in non-compatible than in compatible contexts in the first syllable given the similarity of the f0 contours. Therefore, it seems more reasonable to assume that children's difficulties in producing non-compatible tone contours are production-based.

The finding that children move the f0 toward the onset f0 of the following tone earlier in the first syllable suggests that children demonstrate more anticipatory coarticulation than adults, a phenomenon reported in other studies that examined segmental speech sound development (Goodell & Studdert-Kennedy, 1993; S. Nittrouer, 1993). However, unlike the anticipatory coarticulation for segment sequences that involve the temporal overlapping of the articulatory gestures for two speech targets that are produced by different articulators, the anticipatory coarticulation of disyllabic tones in children seems more likely to involve an earlier truncation of the articulatory command

for the tone in S1 by greater temporal overlapping of the articulatory command for the tone in S2, than for adult productions. Production of a rising f0 contour mainly involves the activation and contraction of the cricothyroid muscle. A falling f0 is mostly produced by decreasing activation of the cricothyroid muscles and possible involvement of the strap muscles to lower the larynx. In consideration of the neuro-muscular mechanism for f0 changes, it is unlikely that a muscle is activated for one tonal target and deactivated for another tonal target at the same time. It is more likely that the second command is executed earlier before the first command is fully realized. For example, in the case of producing T21 for T31, the child times the contraction of cricothyroid to approach the high onset f0 for T1 earlier relative to the supralaryngeal gestures (i.e., consonant constriction) for the next syllable, thus producing the low to high transition in the first syllable. Consequently, the adult listeners perceive a rising T2 instead of the low T3 in the first syllable.

As presented above, children's incorrect disyllabic productions involved flattened f0 slopes, less rapid change in f0 direction, and fewer deflections (changes in direction) in the f0 contours, mostly in the first syllable but also in the second syllable. The findings that the f0 contours of the correct disyllabic tones produced by children also showed reduced f0 change in both syllables, although to a much lesser degree, suggested that children did not make categorical changes in tone production but approached the correct adult forms progressively with time. This developmental trend fits the "global-to-specific" account of speech sound development (Goffman & Smith, 1999). According to this view, young children produce less differentiated and less specific articulatory gestures. As children get older, they gradually refine and reorganize their gestures,

producing more differentiated articulatory movements (Goodell & Studdert-Kennedy, 1993; S. Nittrouer, Studdert-Kennedy, & McGowan, 1989).

The present results also showed that when children produced a disyllabic tone combination incorrectly, they tended to produce the contour such that one of the two tones was heard as correct by the adult listeners. This could suggest that children organized their articulatory gestures in the time frame of a syllable and used the syllable as a tone production unit.

However, the pattern of anticipatory coarticulation described above for non-compatible tone combinations is also consistent with the interpretation that young children organize their tone gestures globally over the entire (disyllabic) word. By this account, they appear to focus on achieving the target f0 contour of the final part of the word (the canonical shape of the S2 tone contour) at the sacrifice of the canonical tone contour of S1. That is, they have learned that the tone contour at the end of the word is the most important.

Given that children produced significantly more errors in only one syllable (either S1 or S2), it is more likely that the production unit for children's tone production is the syllable; however, the possibility for the word to be the production unit cannot be totally ruled out. More extensive acoustic analysis of children's correct and incorrect DT productions will be helpful in determining the production unit for children's DT productions.

In general, children seemed to target at having at least one tone correct in producing DTs. Focusing on the correct production of one of the two tones in disyllables could be a developmental strategy children use to cope with the physiological constraints

of their immature laryngeal system. It could also be related to the patterns adopted by adults when they modify disyllable words in "baby talk". In Mandarin baby talk, adults and children tend to duplicate one of the two syllables. For examples, <shua1ya2> 'brush teeth' is sometimes produced as <shua1ya2ya2>; <mian4tiao2> 'noodles' is produced as 'mian4mian4'; and <ji1dan4> 'egg' becomes <dan4dan4> in baby talk. However, not all adults use baby talk when they talk to young children.

The finding that children tended to produce the tone in the second syllable more correctly than in the first syllable suggests that children pay more attention to the tone in the second syllable or in utterance final position. It could also be that the second syllable is longer in duration so children are more likely to reach the target in the time frame of the syllable. However, given that no perceptual data was available and the current study was not designed to examine syllable duration and involved words with different syllable structures, no conclusive information can be provided in terms of duration differences. All these speculations need to be confirmed or rejected by future studies.

In summary, children had more difficulty producing more complex f0 contours in non-compatible disyllabic words and we hypothesize that their difficulties are motoric in nature. When they made tone errors in disyllabic words, they tended to produce one of the tones, usually the second tone, correctly. When they produced disyllabic tones, they tended to produce f0 contours with slower f0 changes and fewer changes in f0 direction. They were more likely to change the f0 contour in the first syllable such that the f0 offset of the first tone was closer to the f0 onset of the second tone, which resulted in smaller f0 differences between the tones, particularly when the two tones were non-compatible. Given that children produce slower articulatory gestures (Goffman & Smith, 1999; Smith,

2006; Smith, 1991; B. Smith, 2006) and have immature speech motor coordination (Smith & Zelaznik, 2004; Smith, 2006; Walsh & Smith, 2002), these findings suggest that children's development in disyllabic tone production may be limited by physiological constraints in children's laryngeal gestures.

## Context Effects on Disyllabic Tone Acquisition

All previous studies of the development of tone production reported children's accuracy rates and order of acquisition of the four Mandarin tones without taking context into consideration. This study found that without taking syllable and tone compatibility effects into consideration, no significant differences were found in the relative accuracy of the four Mandarin tones. This was true for the first syllable and when both the first and second syllables were combined. In the second syllable in non-compatible contexts the order of accuracy was T1 ≈ T4 > T2 ≈ T3 and when both the compatible and non-compatible contexts were combined, T1 > T3 ≈ T4 > T2. No difference was found in the accuracy rates among the four tones in the second syllable in compatible context. Thus, the relative accuracy of the four tones differed across contexts; overall, T1 appeared to be easier than the other three contexts, while T2 tended to be the most difficult.

Wong et al. (2005) reported that children produced T1, T2 and T4 with comparable accuracy rates (70% - 78%) in monosyllabic familiar words, whereas T3 was produced significantly more poorly (44%). To compare children in the similar age range, the accuracy rates of the four tones for the three-year-olds in the present study were computed. In non-compatible contexts in the second syllable the order was: T1 ≈ T4 > T3 > T2. The four tones did not differ significantly in accuracy in any other syllable position or compatibility contexts. Thus, there is a difference in the order of acquisition of the

tones in monosyllables versus disyllables. Acoustic analysis of the four tones in monosyllables and disyllables will provide more information on the error patterns of the four tones in monosyllabic versus disyllabic tones.

When syllable position and tone compatibility were taken into consideration, children's accuracy rates of the four Mandarin tones were context dependent (Figure 9). T1 was predominantly affected by syllable position; accuracy rates were significantly lower in the first than in the second syllable in both compatible and non-compatible disyllables. T2 was largely influenced by the compatibility of the disyllabic tone combinations. The accuracy rates of T2 in both the first and second syllable were significantly lower for non-compatible combinations than for compatible disyllables. Children had more difficulties with T3 and T4 in the first syllable in non-compatible combinations. In non-compatible contexts, significantly more errors were found in the first than in the second syllable.

These context effects appear to be related to f0 complexity. In general, children had more difficulties producing the four tones in non-compatible contexts. With T2, the error pattern is obvious. The accuracy rates for T2 were lower in non-compatible contexts than in compatible contexts in both syllable positions. For T3 and T4, children also made more tone errors in non-compatible disyllable combinations, but the error was found primarily in the first syllable, indicating that they changed the f0 contour in the first syllable to reduce the challenges in producing rapidly changing f0 (see discussions above). In the case of T1, children seemed to have difficulty maintaining a relatively level tone in the first syllable in compatible as well as non-compatible tones. Results of acoustic analysis showed that children sometimes produced a rising or a falling contour

instead of a high flat f0 contour for T1 in the first syllable (Figures 10B2, 10B3, 11D & 11E). Sometimes children produced T1 in the first syllable with a flat f0 contour at a lower f0 value (Figure 11J). These findings may be indicative of difficulties reaching the exact f0 height for T1 at the beginning of a production. Follow-up acoustic analysis on children's incorrect productions of T1 in the first syllable in this study, and future physiological studies examining children's accuracy and efficiency in achieving a high f0 target and maintaining a steady and consistent high f0 can provide more information about this phenomenon.

To summarize, the order of accuracy of the production of the four Mandarin tones by young children appears to depend on the syllabic and coarticulatory context. Tone accuracy rates from one context may not predict accuracy rates in other contexts. For example, the order of acquisition of monosyllabic tones is different from the order of the four tones in any context in disyllabic words. The accuracy rates of the same tone varied substantially in different contexts. In addition, there seemed to be some lexical effects on children's accuracy rates of disyllabic tones.

### Children's Language/Educational Experience and Disyllabic Tone Production Accuracy

Our analyses showed that children's disyllabic tone production accuracy was related to the length of time the child attended a Chinese school and was not related to their experience in English schools, their Chinese or English language scores, or their experience in a Mandarin-speaking country. However, these results should be interpreted with caution. First, the length of time children attended a Chinese school was also correlated significantly with chronological age, which confounds the interpretation of the results. Second, these results by no means support a conclusion that demographic

background variables present in this sample of children had no impact on their disyllabic tone development. The reasons are, firstly, the language tools adopted in this study were not good measures of the children's language skills because the tests were designed for very different populations. Second, the child participants in this study were not selected from the full range of Mandarin learning children in the Tri-State area but were children who fitted our language criteria: good Chinese skills, limited exposure to English, family members spoke only Mandarin to the children, and having no exposure to other dialects. An effort was made to select children with little exposure to bilingual or English education. Thus, the child participants are not representative of all children who go to English schools or reside in a country where Mandarin is the dominant language. Thus, the results of the correlation analyses should not be generalized to other populations. The main purpose of carrying out these correlation analyses was to ensure that the performance of the children who were included in this study were not confounded by their language and educational backgrounds.

Given that all the words selected in this study were the most familiar words for the children in the two rounds of word familiarity testing prior to this study, the results that no significant difference was found in the accuracy rates of the disyllabic tones and the number of children who produced the words does not mean that word familiarity had no bearing on tone accuracy.

## Future Studies

Clearly, more extensive research with larger sample sizes is required to confirm and extend the findings of the present study. Studies with larger sample sizes will allow more precise analyses of age differences and confirm the phenomena suggested in this

study such as the regression of performance in 4-year-old children. Detailed acoustic analysis of children's accurate and inaccurate tone productions will be very informative. It will provide more information on children's articulatory gestures while producing the tones in different contexts; it will allow us to track the f0 changes with age and reveal how children approach the adult forms over time. Longitudinal studies would also be helpful in tracking the developmental changes of tone production in the same child speaker and can confirm whether individual four-year-old children experience tone production regression during their development. Perceptual studies on tone may help identify factors that contribute to children's acquisition of lexical tone contrasts.

Examination of children's disyllabic productions in different contexts (e.g., in different lexical items, segmental constructions, utterance positions, length of utterance) could reveal more information on the factors that affect tone production accuracy. Studies that examine variability in tone development (intra-subject and intra-subject or trial-to-trial variability) and the change of variability over time will inform us about the characteristics of the speech production system in children. Results of these studies will help to establish baselines for clinical evaluation and facilitate clinical treatment for children with tone production difficulties and motor speech disorders, and will also inform theories of phonological development.

Studies on the anatomical and physiological development of speech motor control are essential to the development of explanations of developmental patterns of tone production. Detailed information about the developmental milestones and changes in the anatomy and physiology of the neuromotor system and the speech production mechanisms could help us make more educated predictions about the factors that affect

children's tone and segmental speech development. Studies that examine the efficiency, stability, speed, and flexibility of laryngeal gestures (e.g., speed of f0 acceleration and deceleration in children, precision in hitting f0 targets with different f0 heights, abilities to maintain a steady high or low f0 contour) will be valuable in our understanding of children's lexical tone development. Xu and Sun (2002) demonstrated physiological limits on the speed of pitch change in adults and claimed that the maximum speed of pitch change is often approached by the speakers during speech production (Xu & Sun, 2002). However, little is known about children's biomechanical limits on the speed of pitch change. In addition, the role of training and practice on the development of speech motor control and tone accuracy will be valuable for making clinical decisions on treating children with lexical tone production difficulties. If the production of certain tone targets or combinations requires anatomical and physiological maturation and cannot be substantially improved by training or practice, therapeutic goals should be designed accordingly. Findings in all these studies will not only better our understanding of children's tone development but will also provide important information on children's phonological development in general, and will help shape phonological theories.

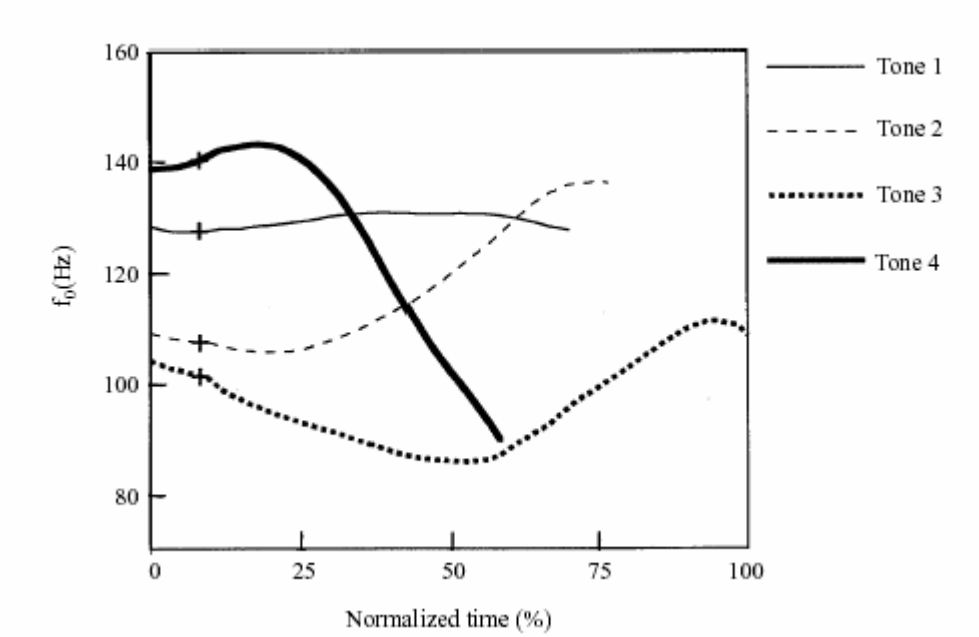
## Conclusions

This is the first study that systematically examines children's production of tones in disyllabic words. It provides preliminary data on children's acquisition of tones including accuracy rates, developmental error patterns and context effects on tone acquisition. Possible underlying causes for children's disyllabic tone production difficulties were discussed.

Overall, the findings suggest that tone development is a protracted process. Even 5- to 6-year-old children do not produce most disyllable tone contours as accurately as adults do. Children between 2 and 6 years old improve in overall tone production accuracy and produce more of the 15 tone combinations with adult-like accuracy. Five- to 6-year-old children show a significant improvement in overall accuracy of tone productions, a greater number of disyllabic tones that are produced with adult-like accuracy, and a decrease in inter-subject variability. However, their overall accuracy rates and accuracy rates in some of the disyllabic tones are still not adult-like.

Children's disyllabic tone production accuracy is related to f0 complexity, possibly due to physiological constraints in their immature laryngeal control system. Non-compatible tones which have more complex f0 and involve more rapid f0 changes are produced less well, even by the oldest children tested here. The patterns of errors as a function of tone compatibility and syllable position argue against the interpretation that perceptual difficulties were a major contributing factor to children's tone production difficulties. The findings that there was target undershoot and reduction of f0 slopes in children's disyllabic productions in judged accurate as well as inaccurate productions suggested that physiological constraints influenced children's disyllabic tone productions. The finding that children's accuracy rates are related to f0 complexity and syllabic position argues that research on lexical tone development in continuous speech contexts is necessary for a more complete understanding of this important aspect of phonological development. In general, the findings support biological models and constraint-based theories of phonological development.

Figure 1. Mean f0 Contours of the Four Mandarin Tones in the Syllable /ma/ Produced in Isolation




Note: The time is normalized, with all tones plotted with their average duration proportional to the average duration of Tone 3. The crosses in the figure mark the boundary of the consonant and vowel in /ma/.
Reprint from "Contextual Tonal Variations in Mandarin" by Xu, 1997, Journal of Phonetics, 25, P. 61-83. Figure 2. P.67, Copyright 1997, with permission from Elisevier.

Figure 2. Mean f0 Contours of the 16 Combinations of the Four Mandarin Tones

Figure 2A. F0 Contours in Disyllabic Words with the Same Tone in the First Syllable

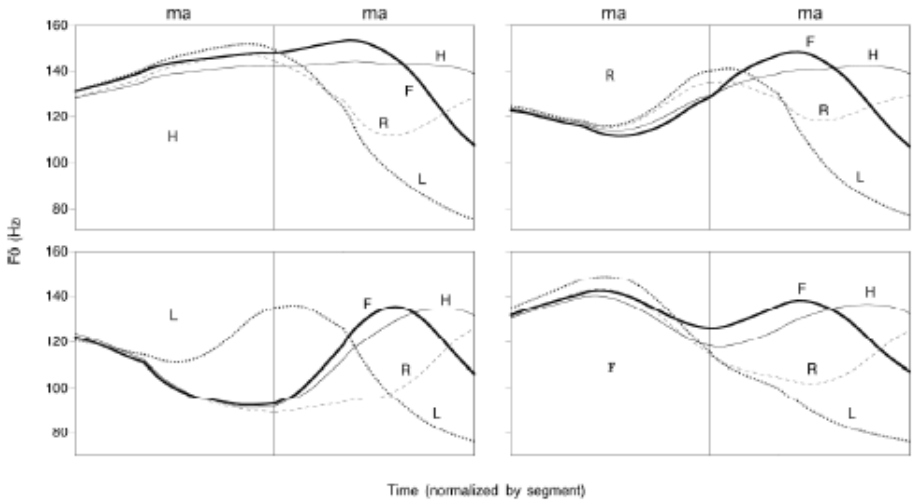
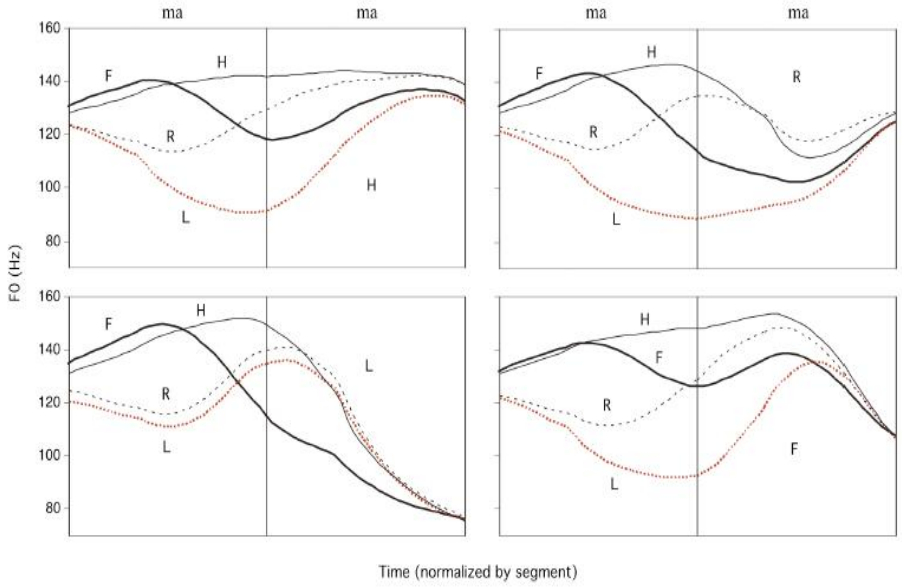



Figure 2B. F0 Contours in Disyllabic Words with the Same Tone in the Second Syllable




Note: Each contour represents 48 tokens of the tone combination produced by each eight native male speakers. The panels in figure 2A are organized by the tone of the first syllable, while the panels in figure 2B are organized by the tone of the second syllable. The vertical lines represent the syllable boundaries (Xu, 2001)

Figure 3. Overall Accuracy Rates of Children's and Adults' DTs by Age Group

Figure 3A

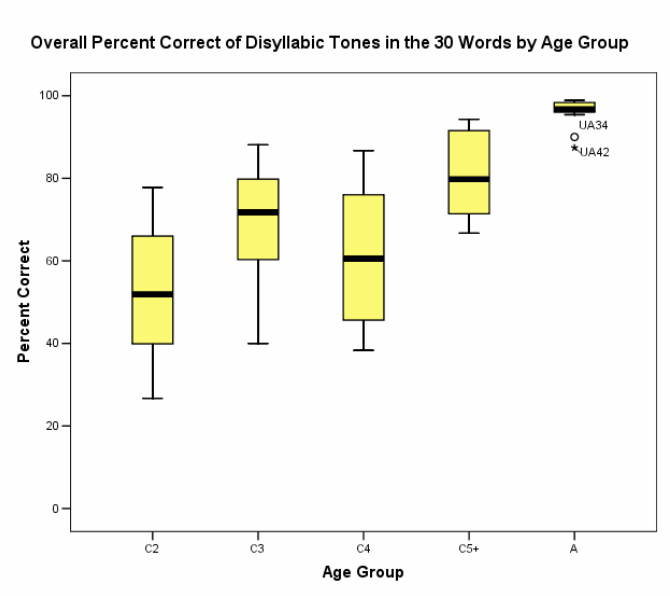
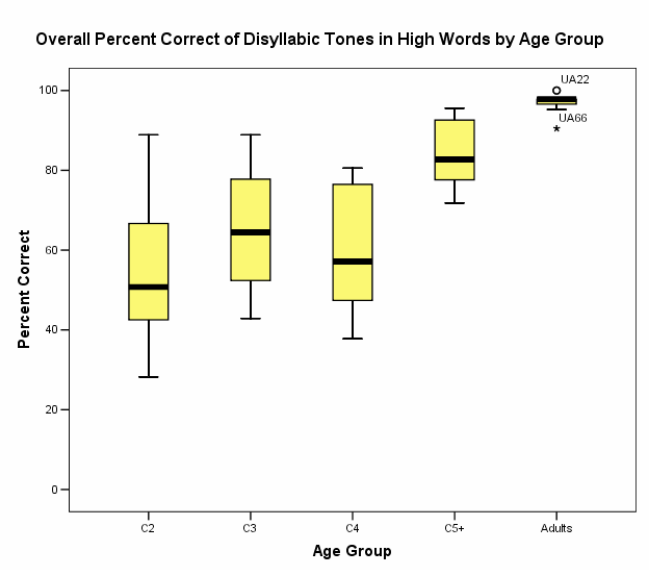



Figure 3B




Note: Cases marked with a circle are outliers (with scores falling between 1.5 to 3 box-lengths from the 75$^{th}$ percentile or 25$^{th}$ percentile).
Cases marked with asterisks are extremes scores (with values falling beyond 3 box-lengths from the 75$^{th}$ percentile or 25$^{th}$ percentile).
Given that adults' scores are mostly at ceiling, with very small interquartile ranges, the outliers and extreme adult scores were included.

Figure 4. Accuracy Rates of the 15 DTs Produced by Children and Adults

Figure 4A

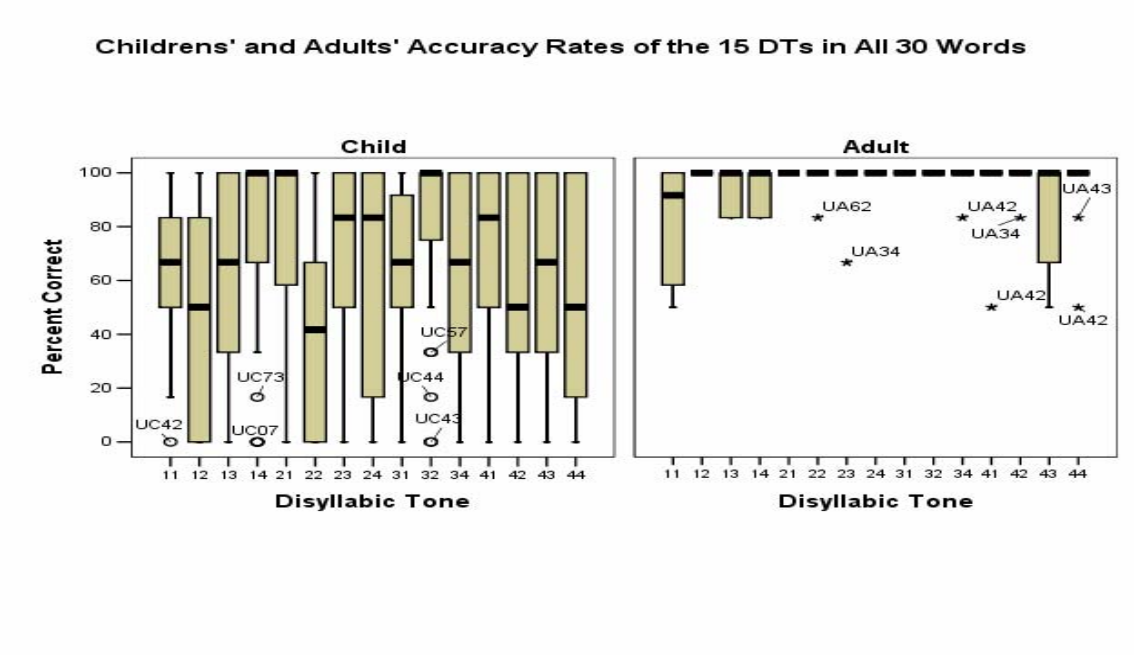



Figure 4B

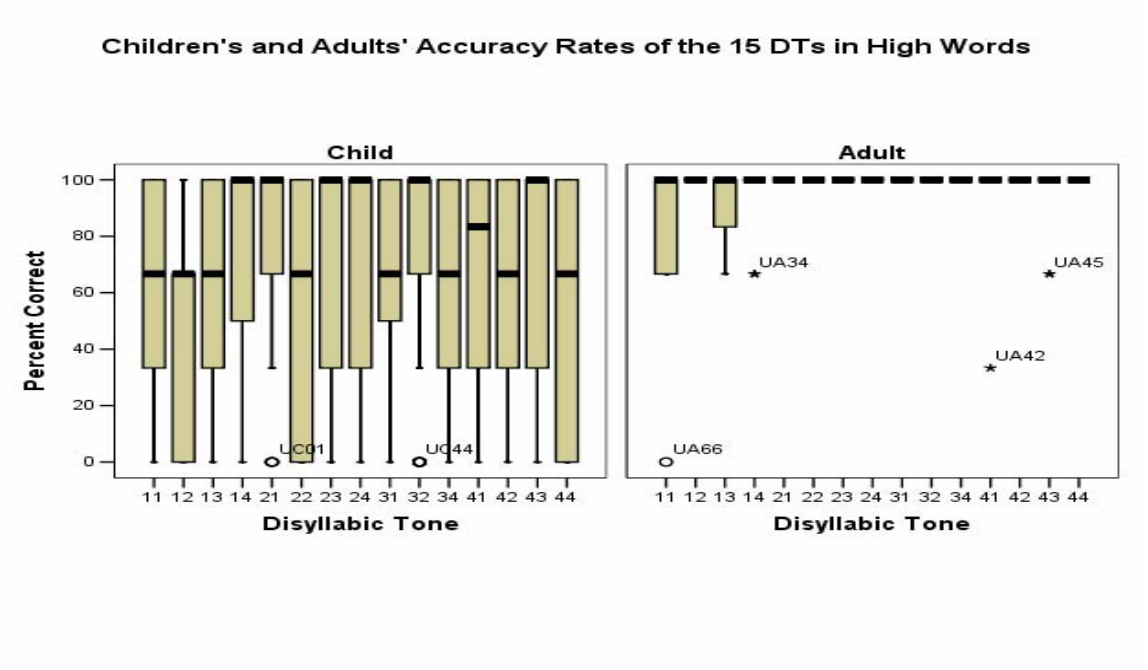


Figure 5. Accuracy Rates of the 15 DTs by Age Group

Figure 5A

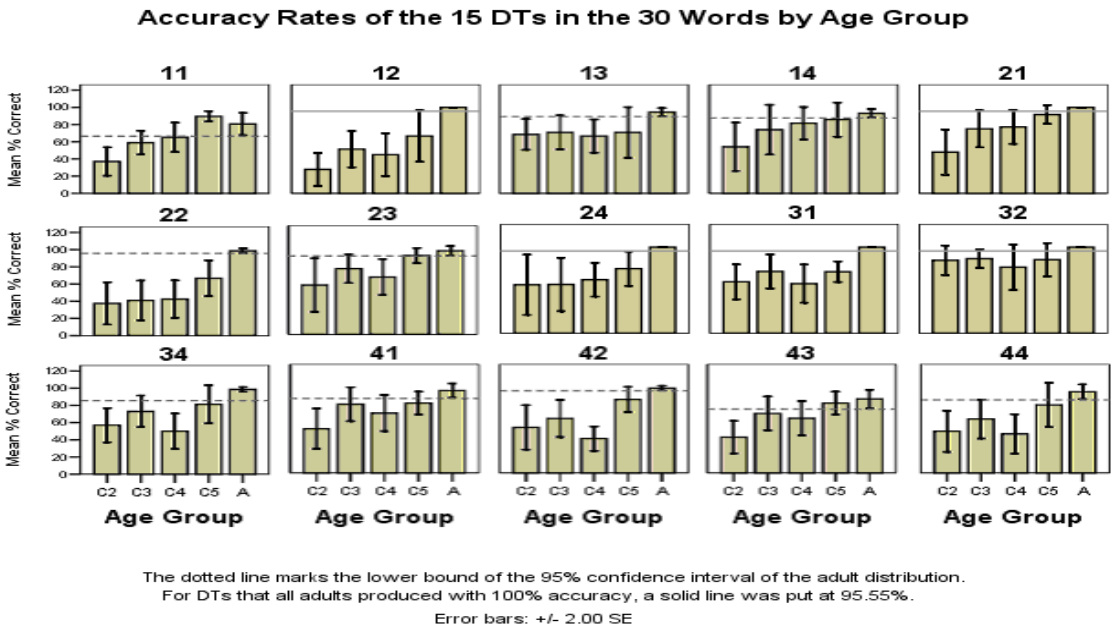



Figure 5B

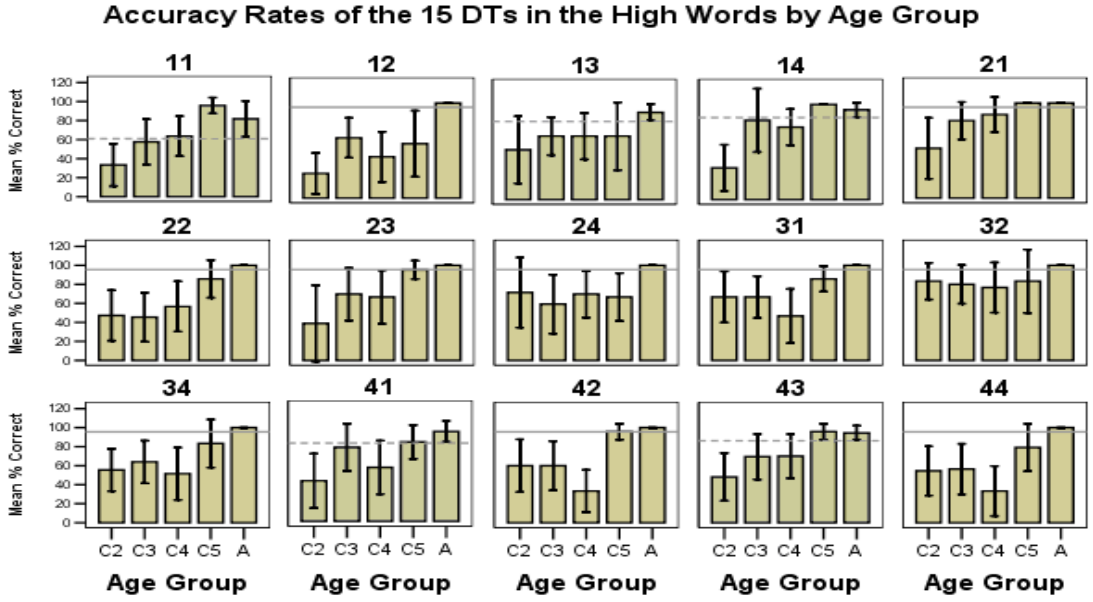


Figure 6. Development of Children's Overall Accuracy in DTs

Figure 6A

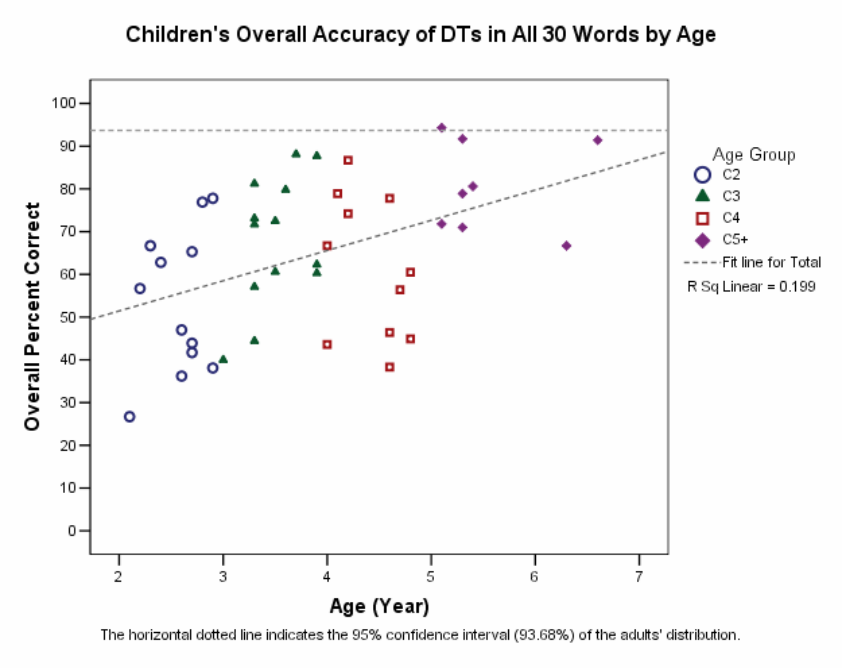



Figure 6B

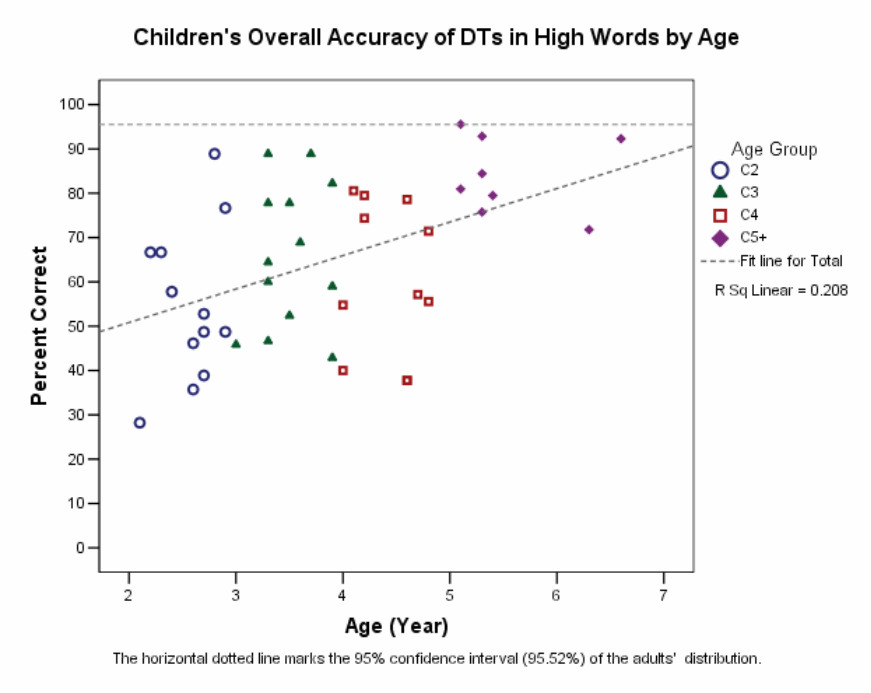
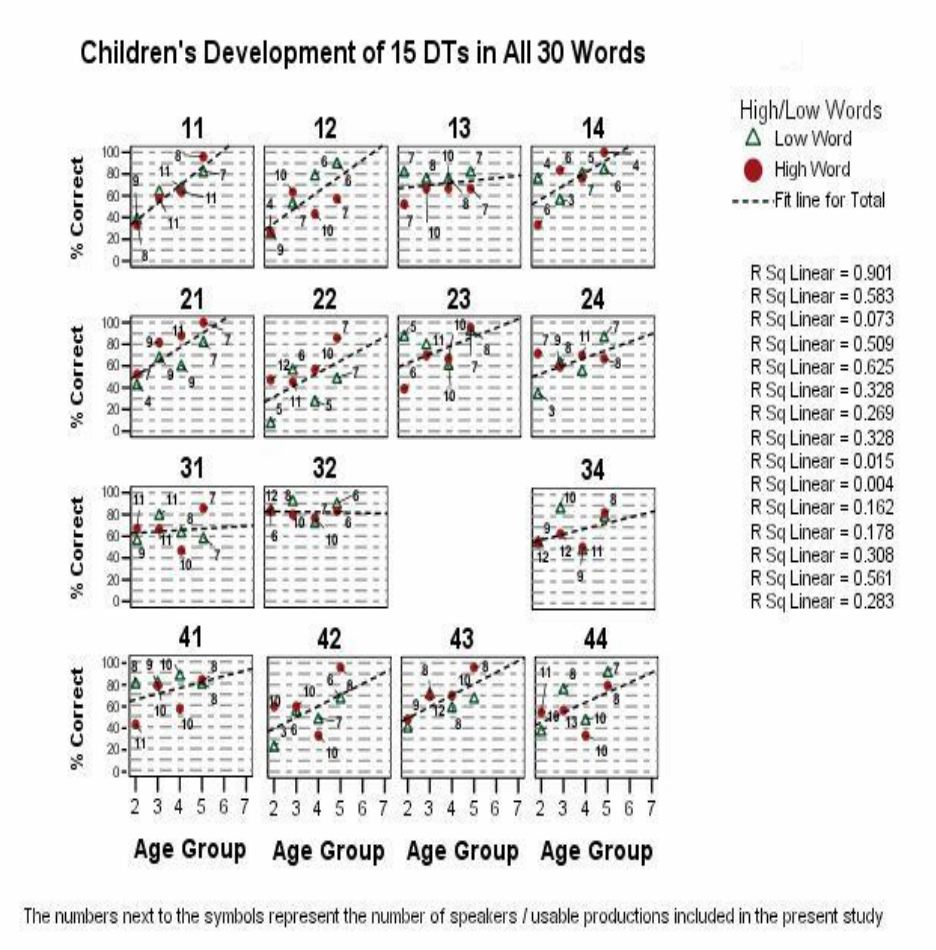


Figure 7. Children's Development of the 15 DTs

Figure 7A




The numbers next to the symbols represent the number of speakers / usable productions included in the present study

Figure 7B

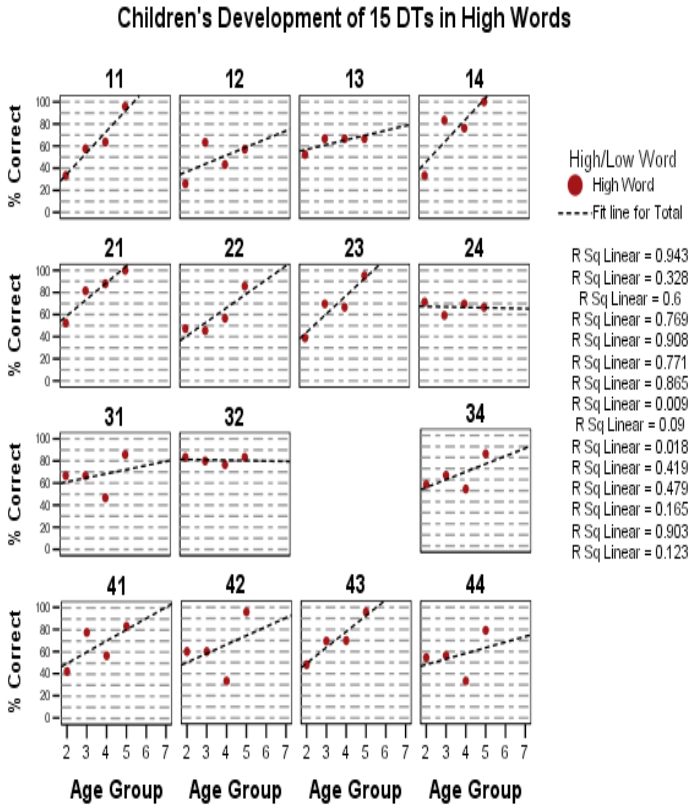


Figure 8. Percent of One- versus Two-syllable Errors in High Words by Children

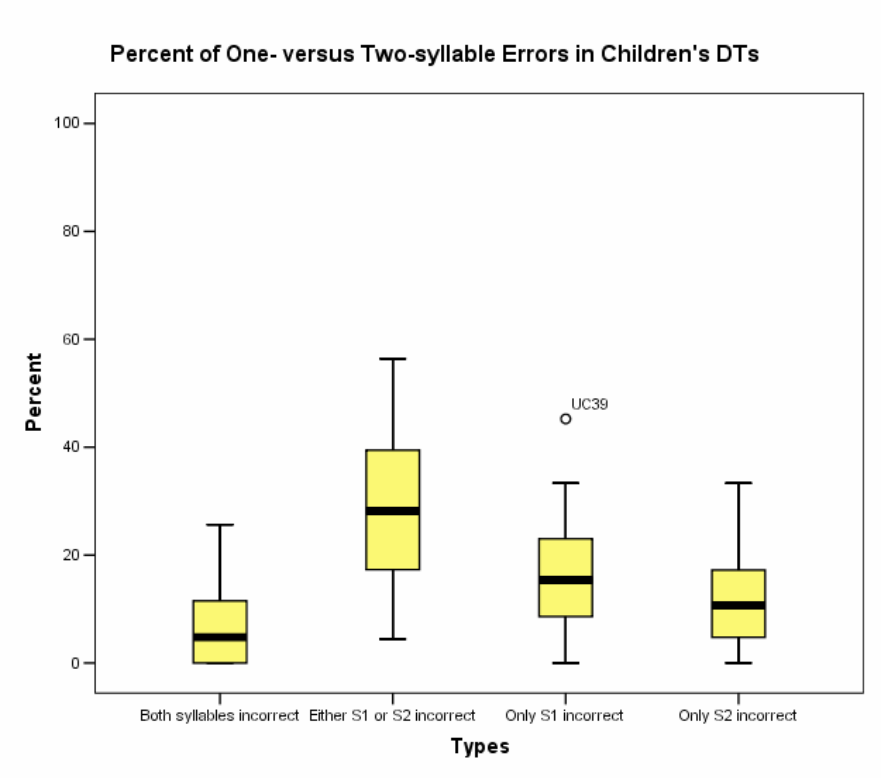


Figure 9. Children's Accuracy Rates of the Four Tones in High Words

Figure 9A

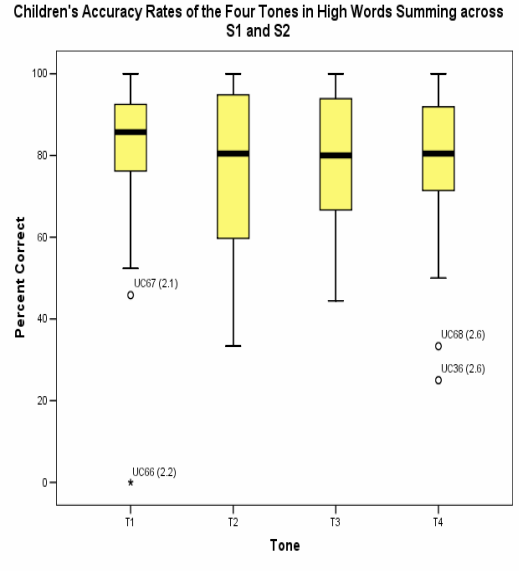



Figure 9B

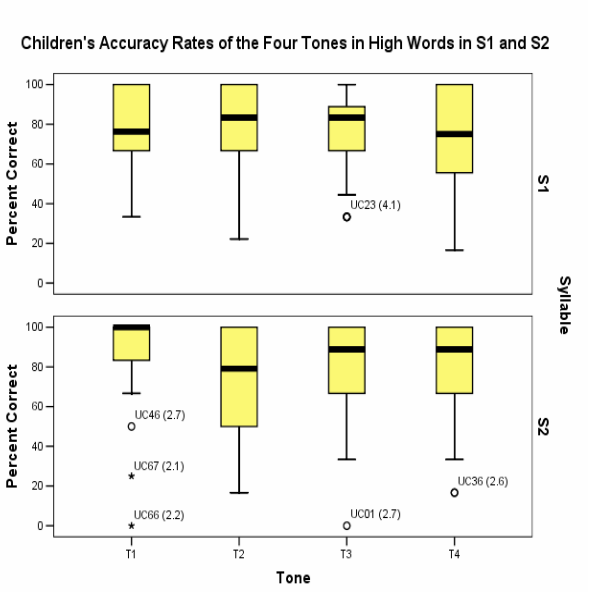
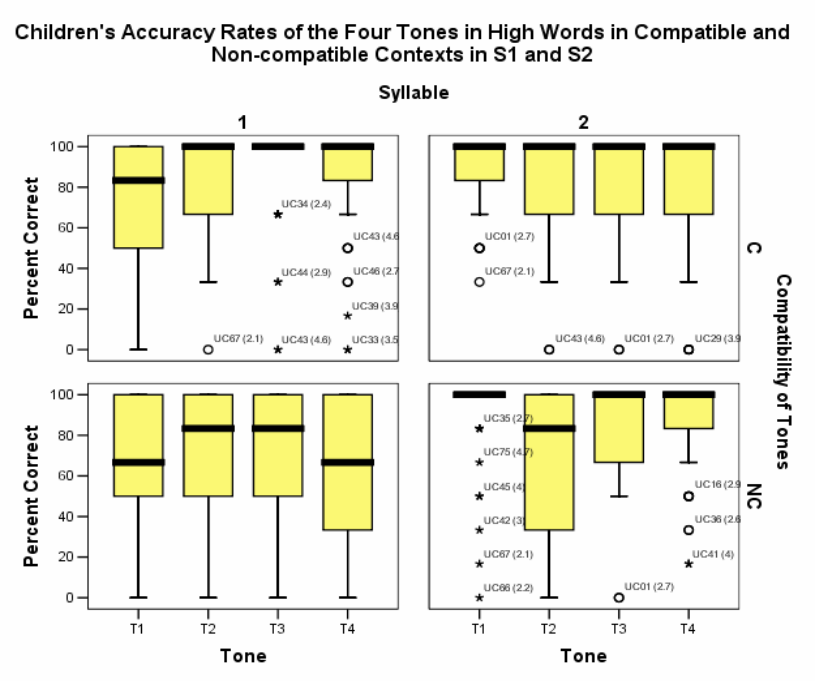



Figure 9C




Note: The numbers in parenthesis represent the age of the child (e.g., 2.1 years old ).

Figure 10. F0 Contours of Correct DT Productions in High Words by Selected Adults and Children

Correct Adult Productions

Figure 10 A1

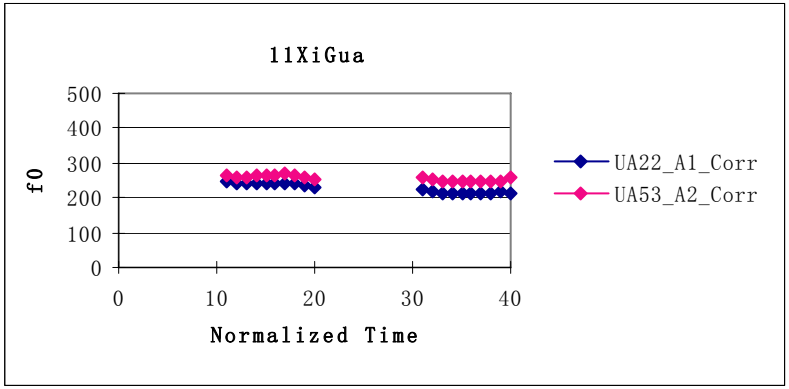
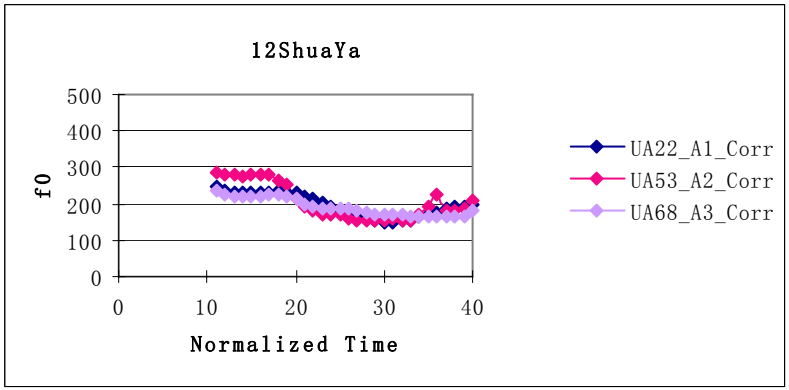



Figure 10 A2




Correct Child Productions [a, b]

Figure 10 B1

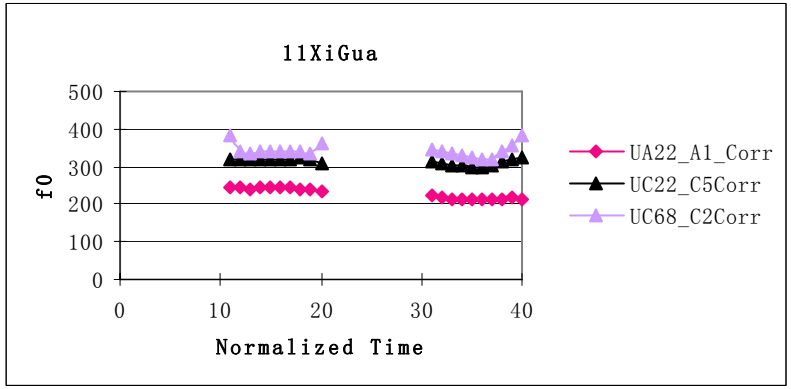
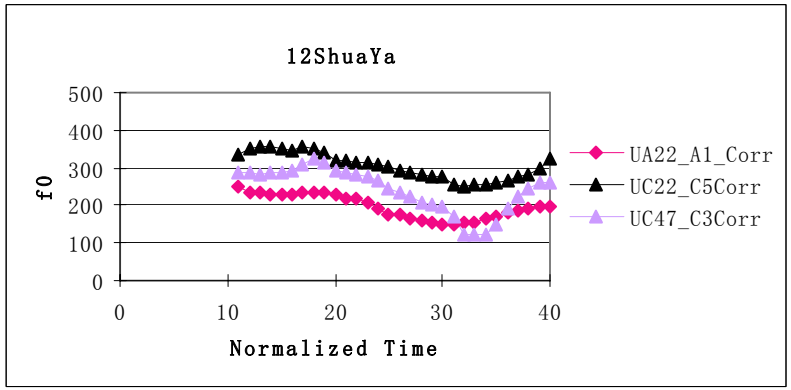



Figure 10 B2



Correct Adult Productions

Figure 10 A3

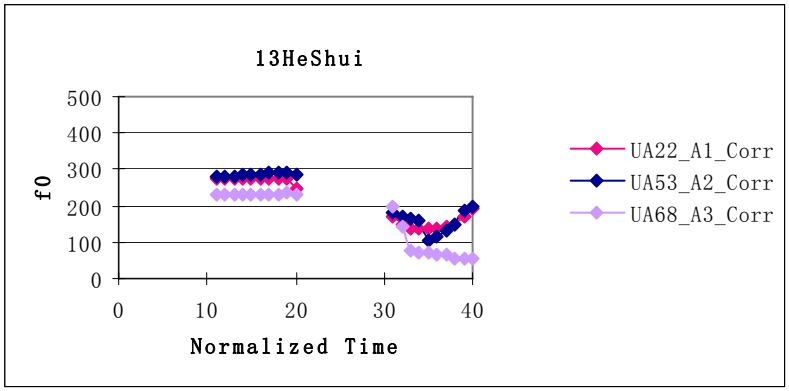
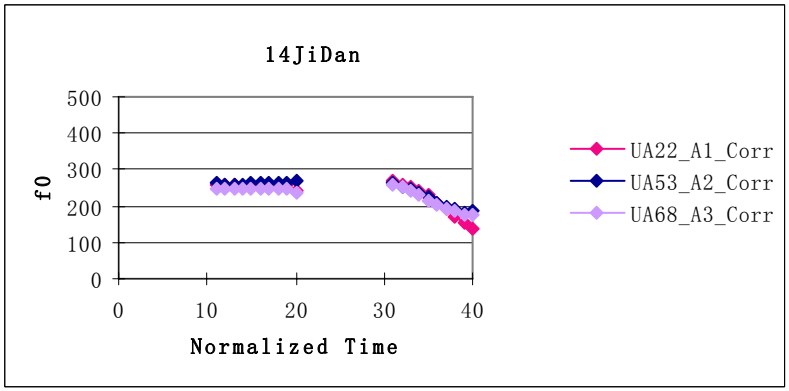



Figure 10 A4




Correct Child Productions [a, b]

Figure 10 B3

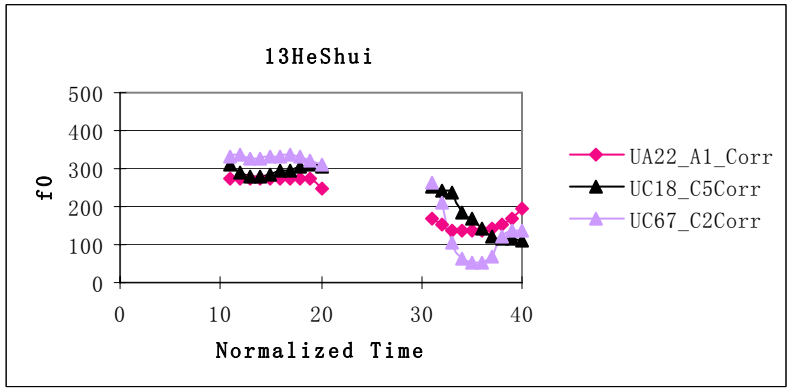
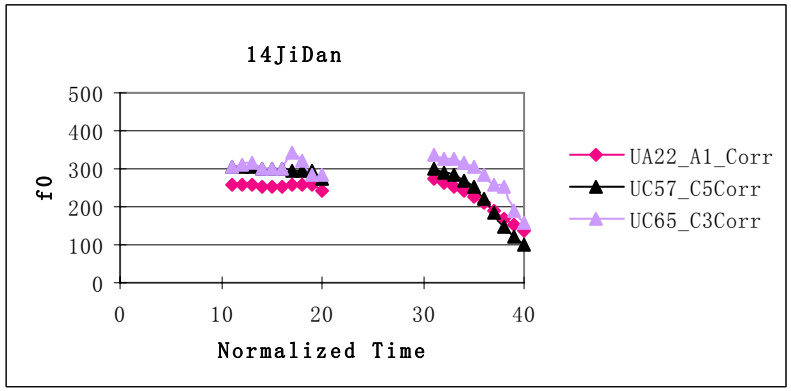



Figure 10 B4



Correct Adult Productions

Figure 10 A5

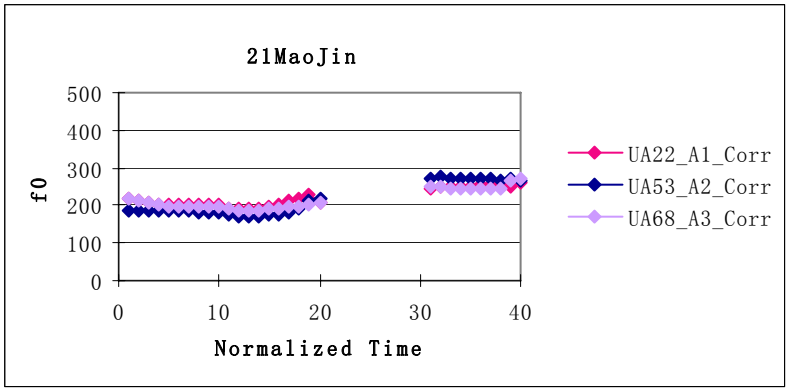
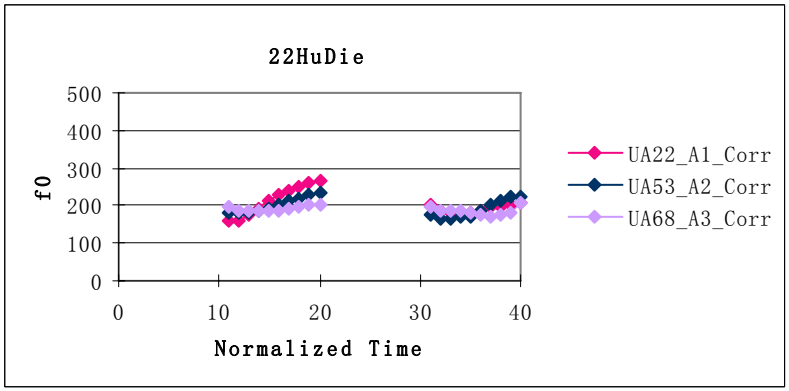



Figure 10 A6




Correct Child Productions [a, b]

Figure 10 B5

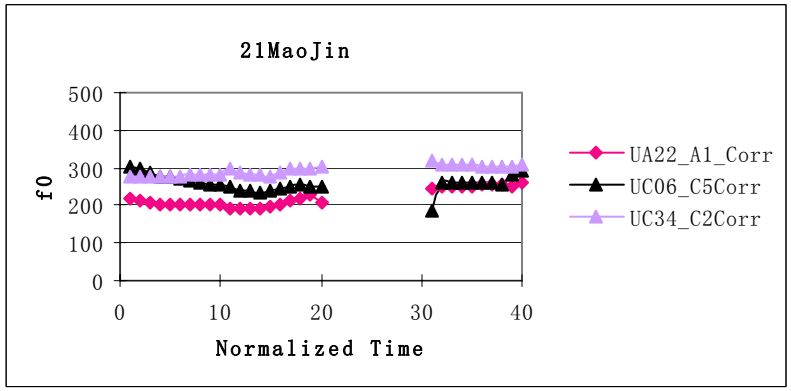
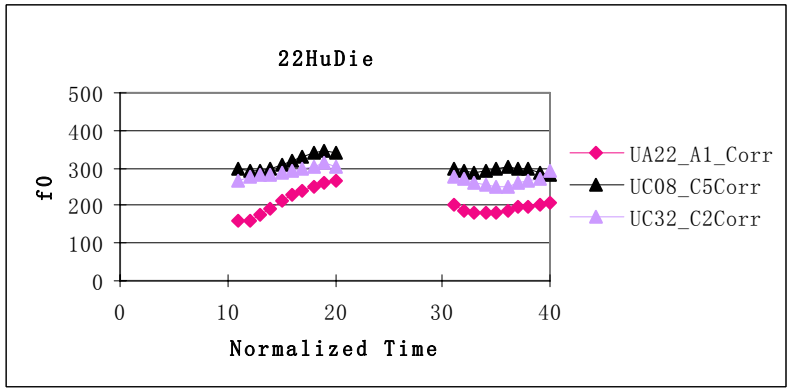



Figure 10 B6



Correct Adult Productions

Figure 10 A7

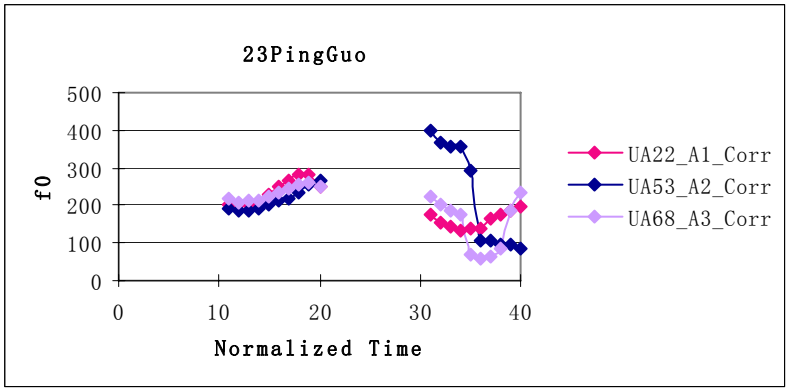
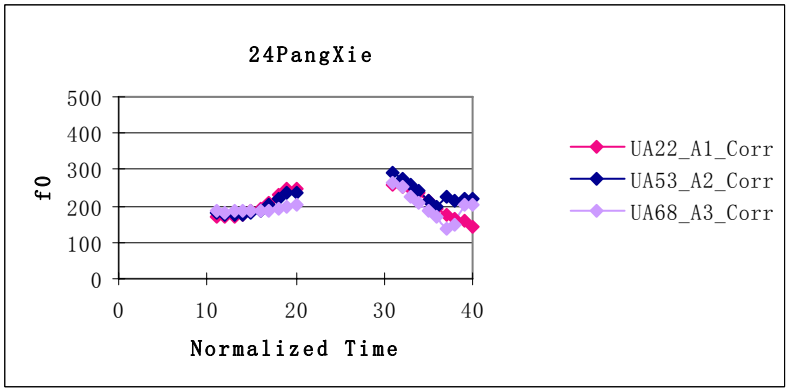



Figure 10 A8




Correct Child Productions [a, b]

Figure 10 B7

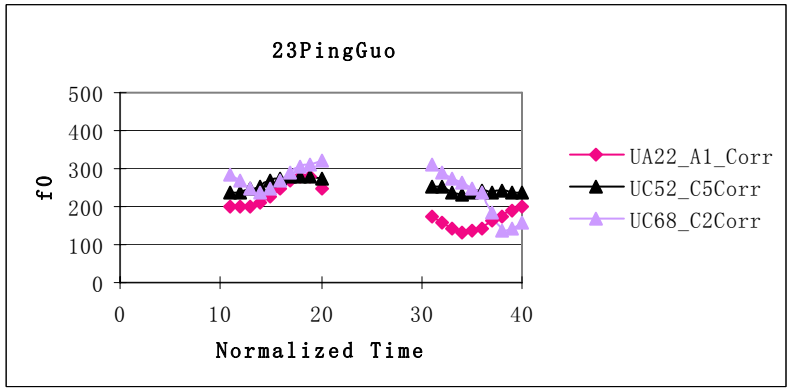
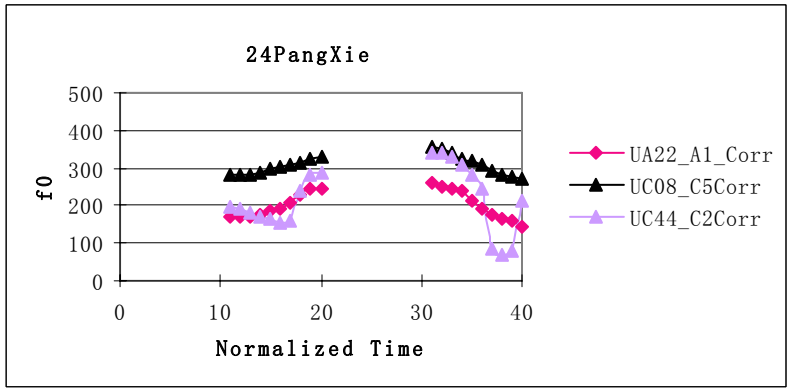



Figure 10 B8



Correct Adult Productions

Figure 10 A9

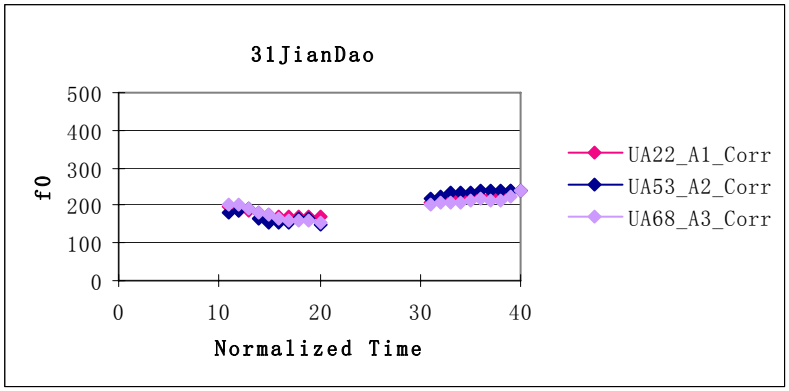
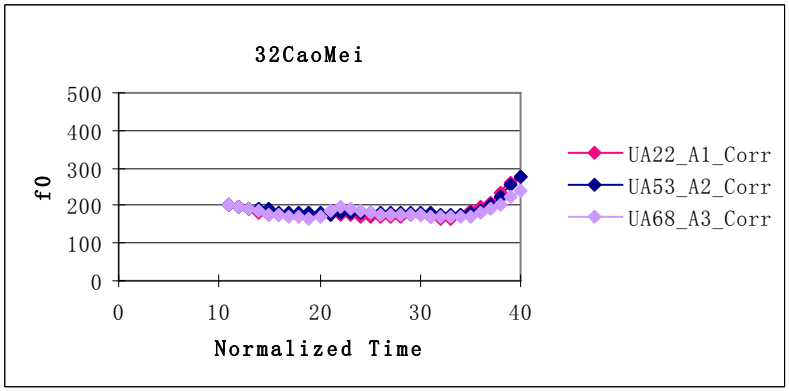



Figure 10 A10




Correct Child Productions [a, b]

Figure 10 B9

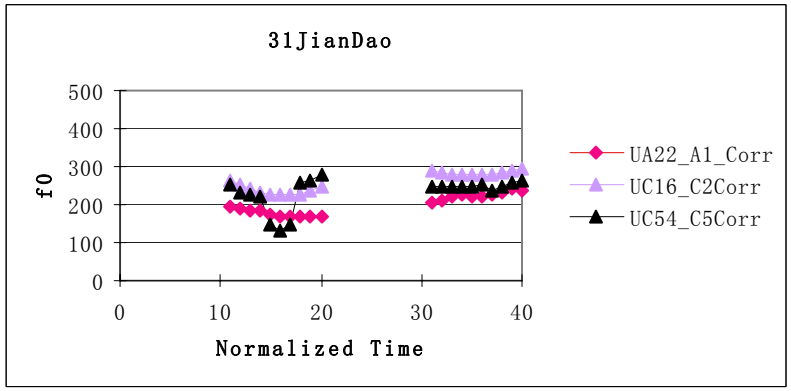
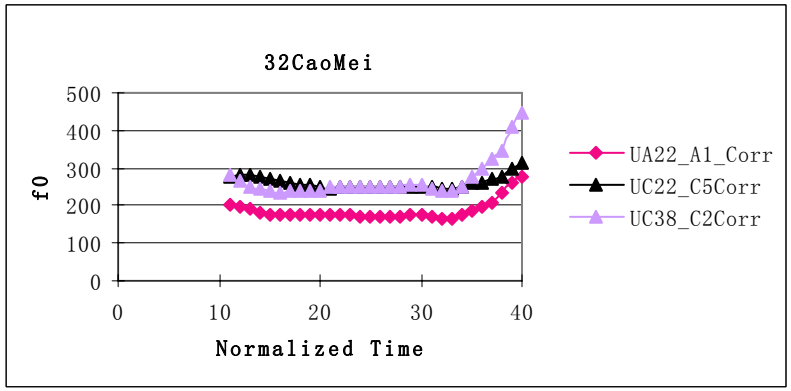



Figure 10 B10



Correct Adult Productions

Figure 10 A11

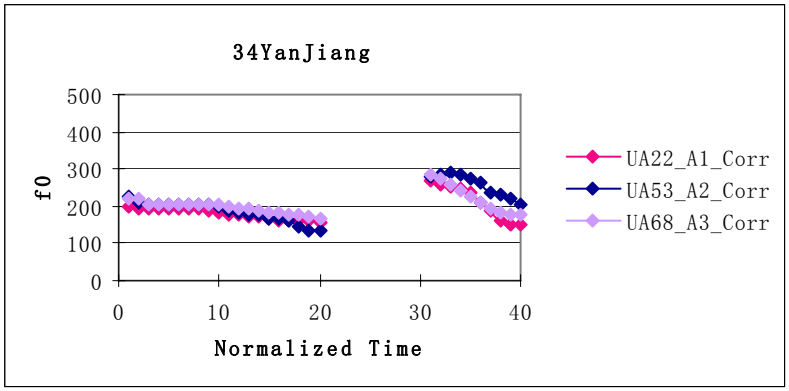
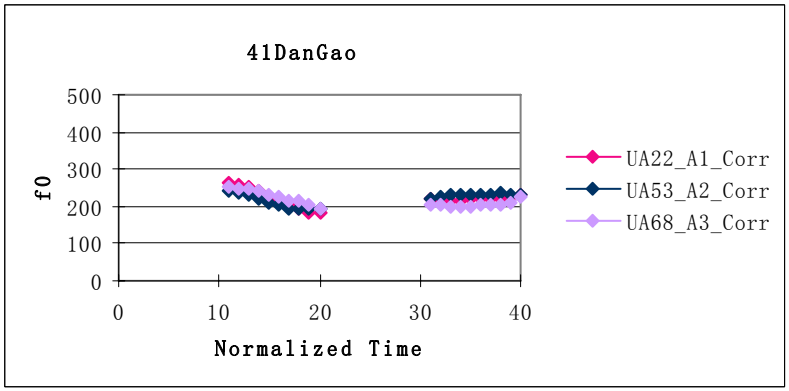



Figure 10 A12




Correct Child Productions [a, b]

Figure 10 B11

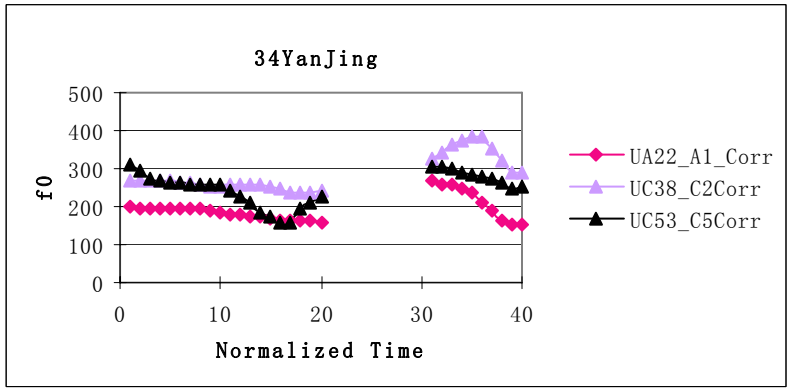
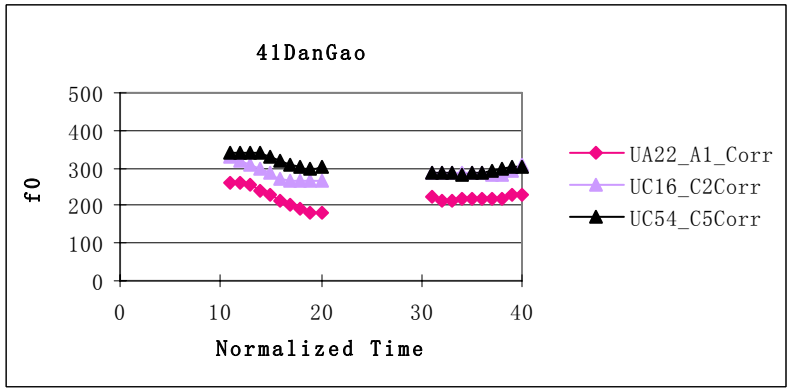



Figure 10 B12



Correct Adult Productions

Figure 10 A13

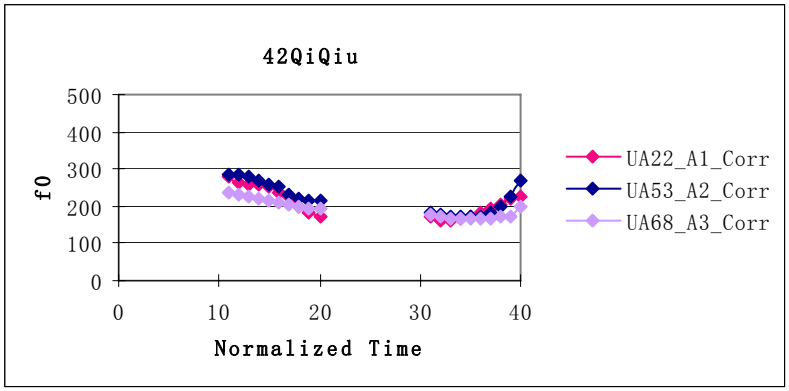
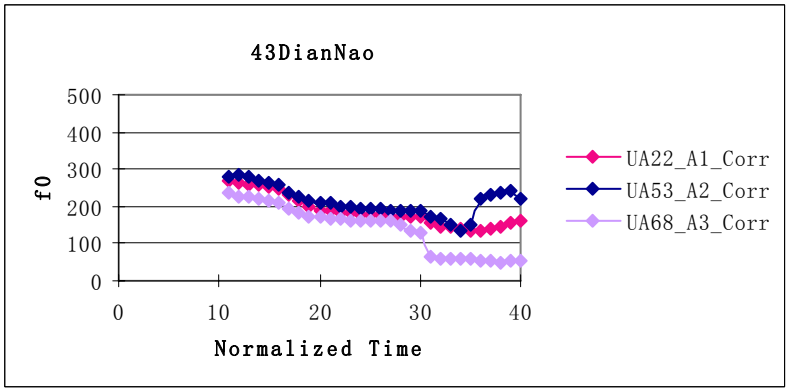



Figure 10 A14




Correct Child Productions [a, b]

Figure 10 B13

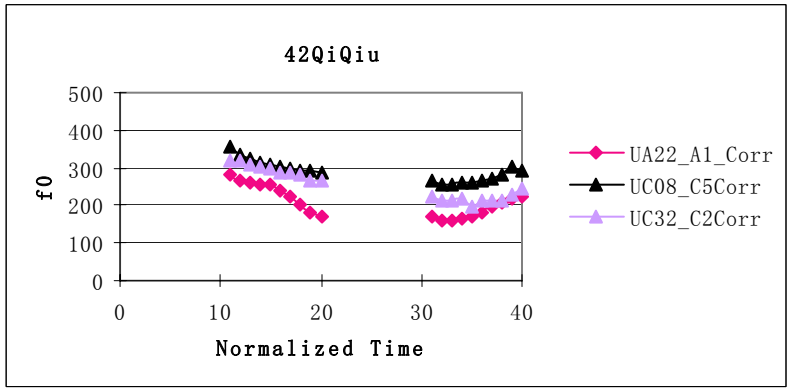
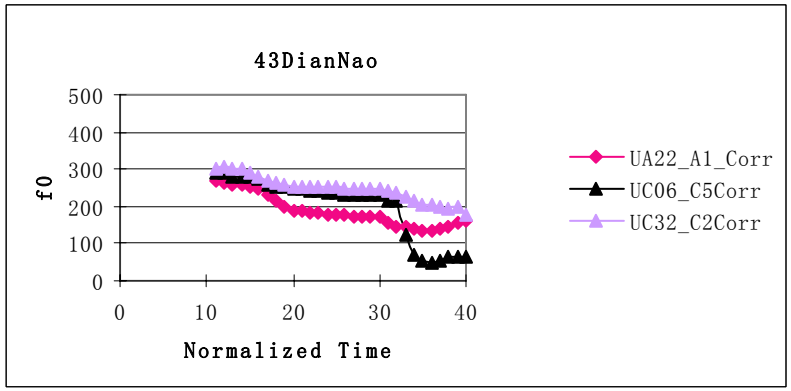
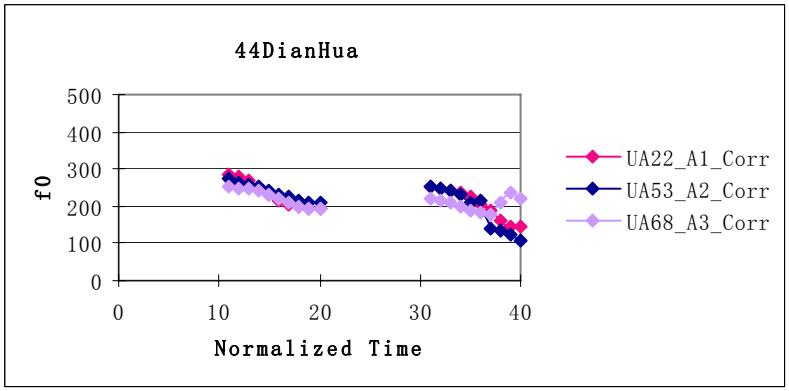



Figure 10 B14



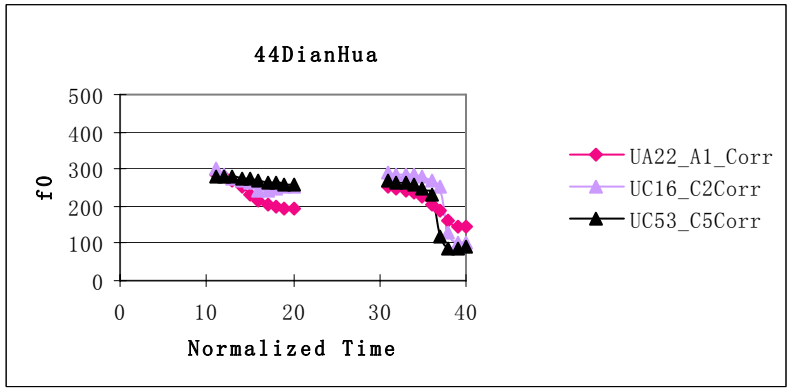
Correct Adult Productions

Figure 10 A15




Correct Child Productions [a, b]

Figure 10 B15

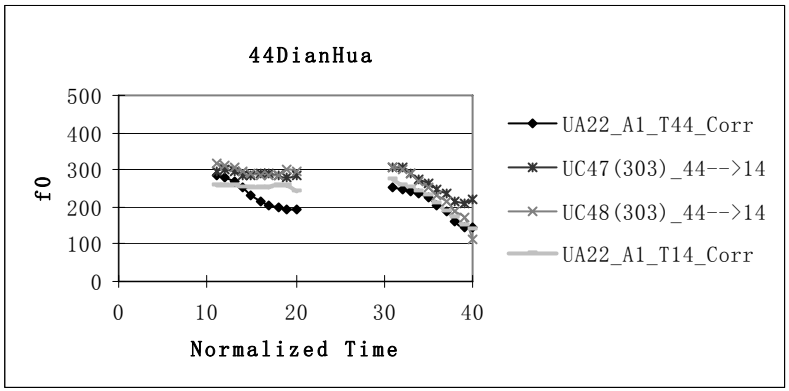



[a] The f0 contour for the same DT correctly produced by an adult, UA22, was included for comparisons.
[b] One correct production from a younger child (children from C2 and C3) and one correct production from an older child (children in C5+) were included.
Note: The X-axis is normalized time intervals. Time points 0-10 represent the initial consonant for S1. Time points 11-20 represent the rime of S1. Time point 20 marks the syllable boundary between S1 and S2. Time points 20-30 represent the initial consonants of S2. Time points 30-40 represent the rime of S2. No f0 contours occur in voiceless consonants.

Figure 11. F0 Contours of Children's Consistent Errors, Adult's Productions of the Target Tones and Adults' Productions of the Substituted Tones

Incorrect Productions by Children [a, b]

Figure 11A. F0 Simplification in S1




Incorrect Productions by Children [a, b]

Figure 11B. F0 Simplification in S1

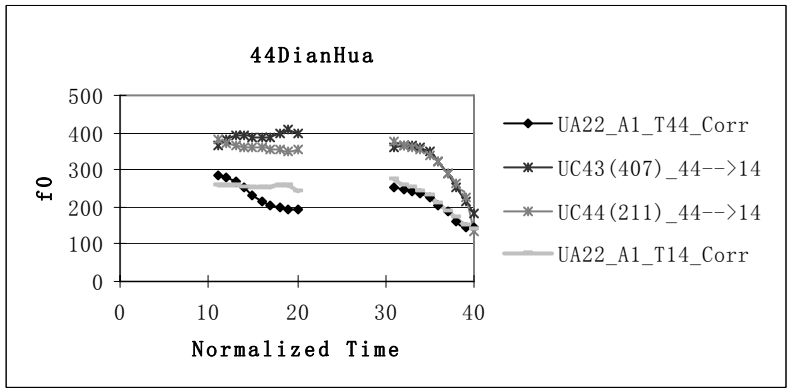



Figure 11C. F0 Simplification in S1

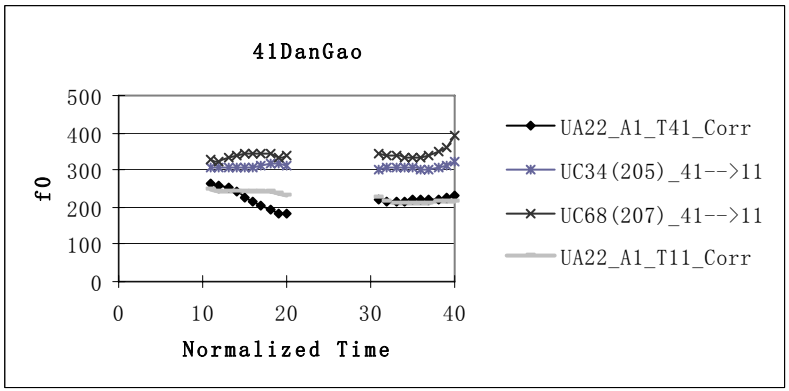
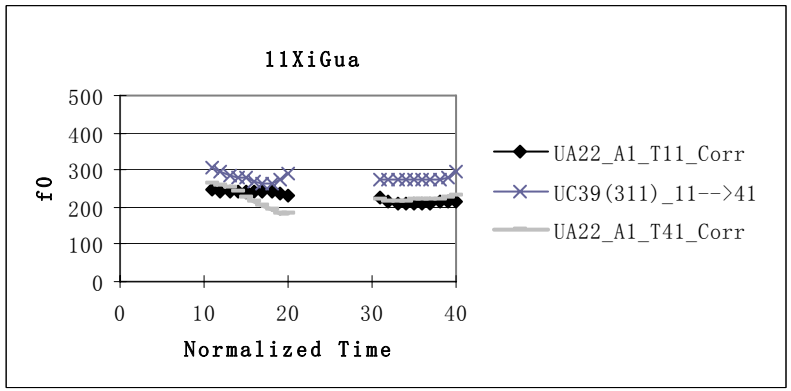



Figure 11D. F0 Simplification in S1



Incorrect Productions by Children [a, b]

Figure 11E. F0 Simplification in S1

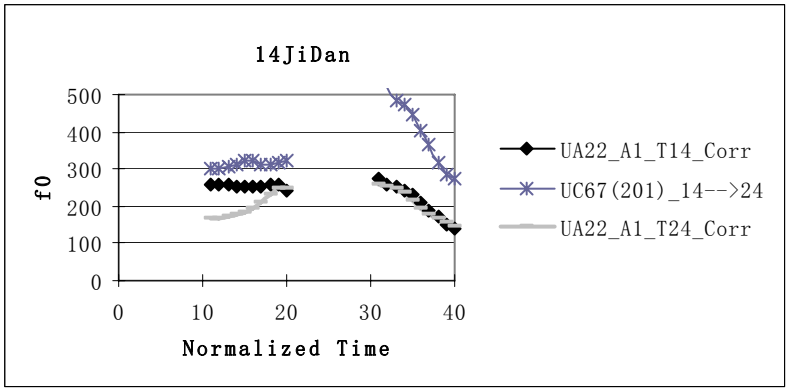
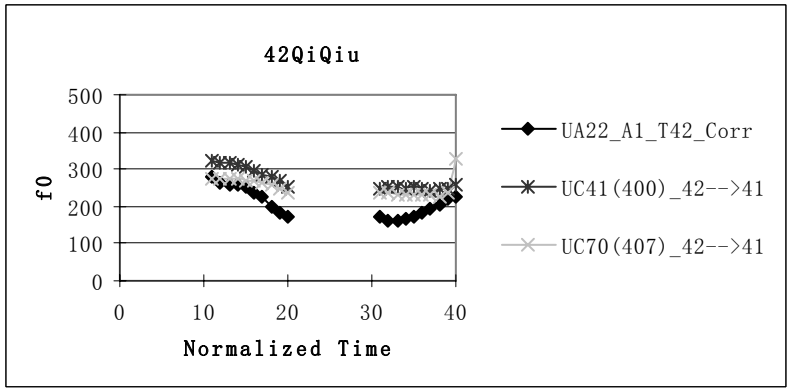



Figure 11G. F0 Simplification in both S1 and S2




Incorrect Productions by Children [a, b]

Figure 11F. F0 Simplification in S2

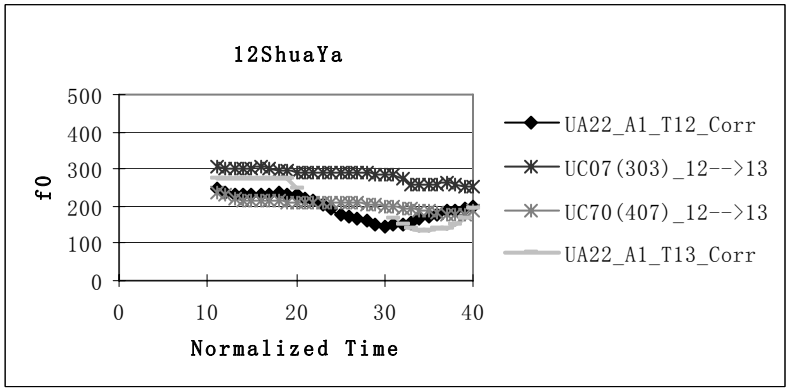
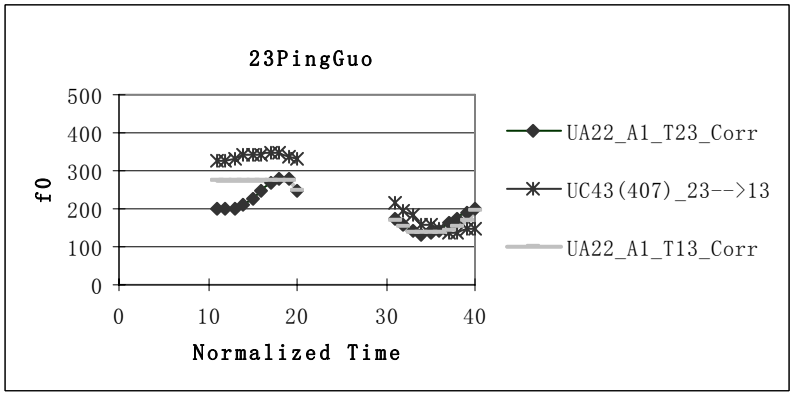
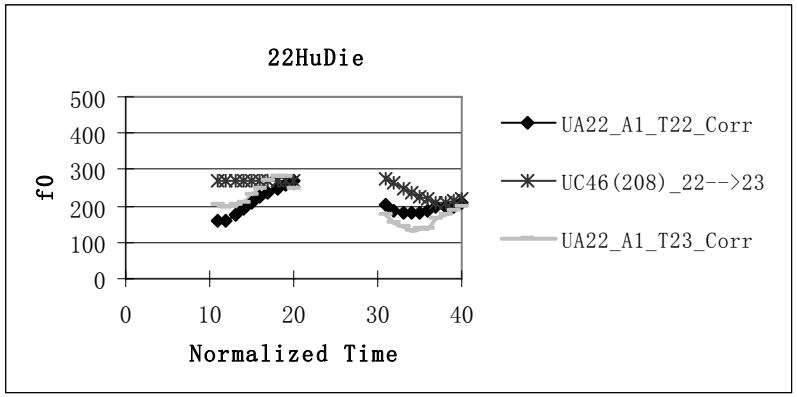



Figure 11H. F0 Simplification in both S1 and S2



Incorrect Productions by Children [a, b]

Figure 11I. F0 Simplification in Both S1 and S2




Incorrect Productions by Children [a, b]

Figure 11J. F0 of T1 Higher in S2 than S1

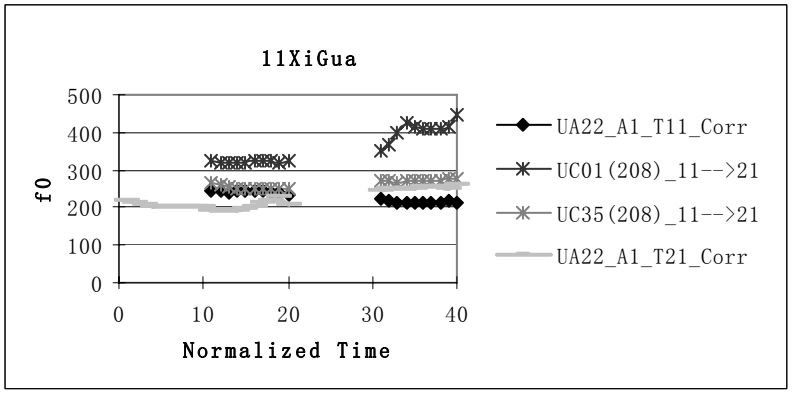



Figure 11K. Different Tone Target in S1

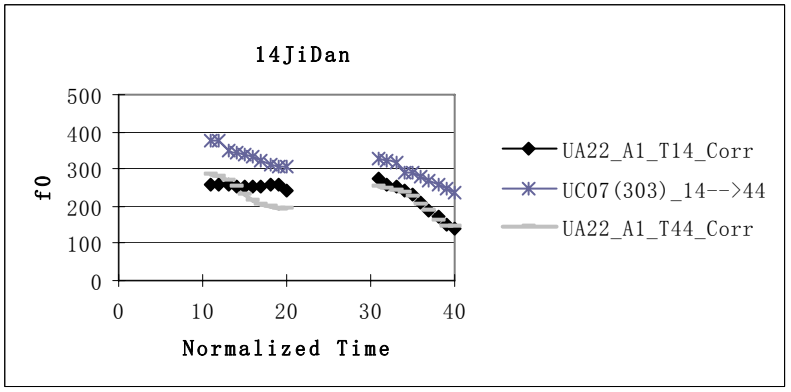
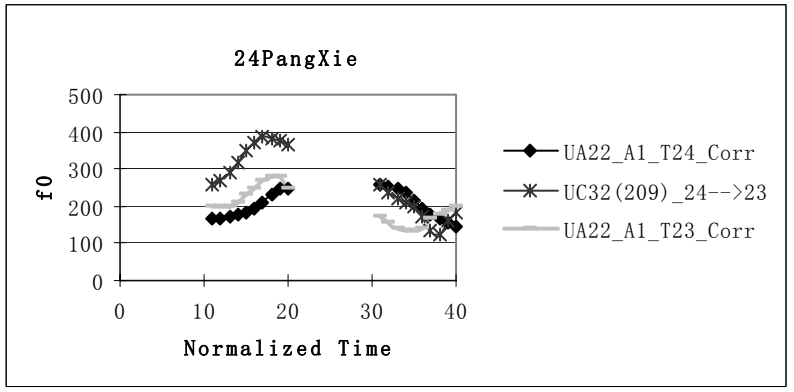
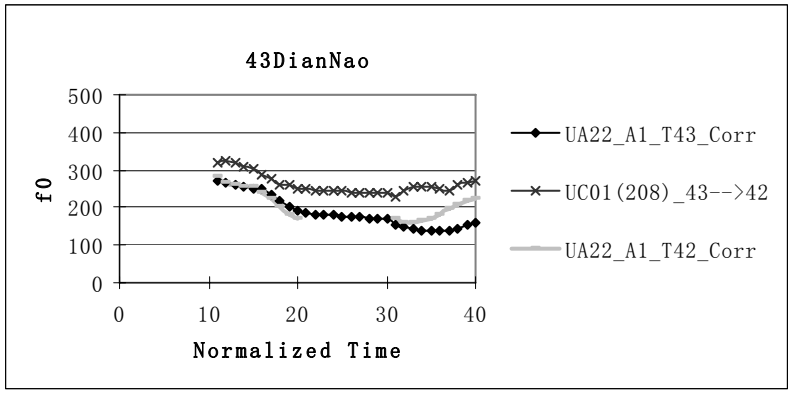



Figure 11L. Different Tone Target in S2



Incorrect Productions by Children [a, b]

Figure 11M. Different Tone Target or Target Undershoot in S2




[a] The child productions were categorized by all three judges into a same incorrect DT.

[b] The f0 contours for the same target DT and the misperceived DT in the adult forms correctly produced by an adult, UA22, were included for comparisons.

Note: The children's ages were presented in the parenthesis in the legend. E.g., (201) represents 2 years and one month old. Substitution patterns were specified in the legend. E.g., 11→21 represents the identified of the target DT T11 as T21 by all 3 judges. The X-axis is normalized time intervals. Time points 0-10 represent the initial consonant for S1. Time points 11-20 represent the rime of S1. Time point 20 marks the syllable boundary between S1 and S2. Time points 20-30 represent the initial consonants of S2. Time points 30-40 represent the rime of S2. No f0 contours occur in voiceless consonants.

Table 1. Compatible and Non-compatible Tone Combinations in Disyllabic Words

| Compatible Tone Combinations (C) | Non-compatible Tone Combinations (NC) |
|---|---|
| T1T1 ( ¯ ¯ ), T1T4 ( ¯ \ ) | T1T2 ( ¯ / ), T1T3 ( ¯ _ ) |
| T2T1 ( / ¯ ), T2T4 ( / \ ) | T2T2 ( / / ), T2T3 ( / _ ) |
| T3T2 ( _ / ) | T3T1 ( _ ¯ ), T3T4 ( _ \ ) |
| T4T2 ( \ / ), T4T3 ( \ _ ) | T4T1 ( \ ¯ ), T4T4 ( \ \ ) |

Note. Because T3T3 (_ _) is produced as T2T3 due to tone 3 sandhi rule, it is not listed in the table. The symbols ( ¯, / , _ , \ ) are schematic representations of the f0 contours for H, R, L, F, respectively.

Table 2. Number of Words Produced by Age Group

| | | Usable Productions Only[a] | | | | | | | All Productions[e] | |
|---|---|---|---|---|---|---|---|---|---|---|
| Age Group | # of Children | Two Words[b] | | One Word | | No Word Produced | | Total[d] | Two Words[f] | |
| C2 | 12 | 82[b] | (45.6[c]) | 70 | (38.9) | 28 | (15.6) | 180 | 96[f] | (53.3[g]) |
| C3 | 13 | 105 | (53.8) | 67 | (34.4) | 23 | (11.8) | 195 | 128 | (65.6) |
| C4 | 11 | 116 | (70.3) | 43 | (26.1) | 6 | (3.6) | 165 | 129 | (78.2) |
| C5+ | 8 | 96 | (80) | 21 | (17.5) | 3 | (2.5) | 120 | 108 | (90) |
| Adult | 12 | 173 | (96.1) | 7 | (3.9) | 0 | (0) | 180 | 180 | (100) |

[a] Usable productions: the productions adopted in tone judgment in this study
[b] The number of two-word productions for the DTs by the same speaker. Only productions that were used in tone judgment were counted.
[c] Percent of cases in which both words of the DTs were produced by the same speaker. Only productions that were used in tone judgment were counted.
[d] Total = 15 tones x # of children in the group
[e] All the target word produced which included the ones that were excluded in tone judgment due to noise, non-isolated production, insufficient intensity or loudness
[f] Number of cases in which both target words for the same DT were produced by the same speaker. Target word productions that were excluded from tone judgment (e.g., non-isolated productions of the target words, noisy productions, and productions that were too loud or too soft) were counted.
[g] Percent of cases both words were produced for the DTs.

Table 3. Percentage of Children in Each Age Group whose Accuracy Rates of the 15 DTs were Adult-like

| | | 11 C | 12 NC | 13 NC | 14 C | 21 C | 22 NC | 23 NC | 24 C | 31 NC | 32 C | 34 NC | 41 NC | 42 C | 43 C | 44 NC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | All 30 Words Included | | | | | | | | |
| A | Adult 95%CI | 66.4 | 95.6 [a] | 89.2 | 87.6 | 95.6 [a] | 95.6 | 91.1 | 95.6 [a] | 95.6 [a] | 95.6 [a] | 85.6 | 86.7 | 95.6 | 74.3 | 85.0 |
| C2 | % within 95%CI [b] | 33 | 0 | 33 | 25 | 13 | 25 | 44 | 50 | 17 | 75 | 17 | 25 | 36 | 10 | 18 |
| C3 | % within 95%CI | 54 | 17 | 42 | 67 | 58 | 18 | 42 | 44 | 46 | 50 | 39 | 50 | 40 | 54 | 39 |
| C4 | % within 95%CI | 64 | 0 | 40 | 56 | 55 | 18 | 46 | 27 | 27 | 70 | 18 | 46 | 0 | 36 | 27 |
| C5 | % within 95%CI | 100 | 43 | 63 | 71 | 75 | 38 | 63 | 38 | 0 | 75 | 63 | 38 | 63 | 63 | 63 |
| All [c] | % within 95%CI | 61 | 13 | 44 | 55 | 51 | 24 | 48 | 39 | 26 | 68 | 32 | 39 | 33 | 40 | 35 |
| | | | | | | | | High Words Only | | | | | | | | |
| A | Adult 95%CI | 60.9 | 95.6[a] | 82.1 | 86.2 | 95.6[a] | 95.6[a] | 95.6[a] | 95.6[a] | 95.6[a] | 95.6[a] | 95.6[a] | 82.2 | 95.6[a] | 86.2 | 95.6[a] |
| C2 | % within 95%CI [b] | 22 | 0 | 43 | 0 | 29 | 33 | 33 | 71 | 55 | 75 | 25 | 36 | 40 | 22 | 36 |
| C3 | % within 95%CI | 55 | 30 | 30 | 83 | 67 | 27 | 64 | 44 | 36 | 60 | 42 | 67 | 40 | 58 | 46 |
| C4 | % within 95%CI | 64 | 20 | 50 | 43 | 82 | 40 | 60 | 55 | 30 | 70 | 36 | 40 | 10 | 50 | 20 |
| C5 | % within 95%CI | 100 | 43 | 57 | 100 | 100 | 71 | 86 | 38 | 57 | 83 | 75 | 63 | 88 | 88 | 63 |
| All [c] | % within 95%CI | 59 | 22 | 44 | 52 | 71 | 40 | 62 | 51 | 44 | 71 | 42 | 50 | 42 | 54 | 40 |

[a] All adults' productions were judged with 100% accuracy. Thus the lower bound of the 95% confidence interval of the adults' scores was set at 95.6%.
[b] Percent of children whose DT accuracy rates were higher than or equal to the lower bound of the 95% confidence interval of the adults' scores.
[c] All children as a group (i.e., from two to six years old)

Table 4. Accuracy Rates and Growth Functions of the 15 DTs

| In the order of the tones in S1 | | | | | In the order of the tones in S2 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| DT | C2 [b] | C5+ [c] | $R^2$ | Effect Size[a] | DT | C2 [b] | C5+ [c] | $R^2$ | Effect Size[a] |
| 11 | 33.3 | 95.8 | 0.943 | Very Large | 11 | 33.3 | 95.8 | 0.943 | Very Large |
| 12 | 25.9 | 57.1 | 0.328 | Medium | 21 | 52.4 | 100 | 0.908 | Very Large |
| 13 | 52.4 | 66.7 | 0.600 | Large | 31 | 66.7 | 85.7 | 0.09 | None |
| 14 | 33.3 | 100 | 0.769 | Very Large | 41 | 42.4 | 83.3 | 0.479 | Medium |
| 21 | 52.4 | 100 | 0.908 | Very Large | 12 | 25.9 | 57.1 | 0.328 | Medium |
| 22 | 47.2 | 85.7 | 0.771 | Very Large | 22 | 47.2 | 85.7 | 0.771 | Very Large |
| 23 | 38.9 | 95.2 | 0.865 | Very Large | 32 | 83.3 | 83.3 | 0.018 | None |
| 24 | 71.4 | 66.7 | 0.009 | None | 42 | 60 | 95.8 | 0.165 | Small |
| 31 | 66.7 | 85.7 | 0.090 | None | 13 | 52.4 | 66.7 | 0.6 | Large |
| 32 | 83.3 | 83.3 | 0.018 | None | 23 | 38.9 | 95.2 | 0.865 | Very Large |
| 34 | 55.6 | 83.3 | 0.419 | Medium | 43 | 48.1 | 95.8 | 0.903 | Very Large |
| 41 | 42.4 | 83.3 | 0.479 | Medium | 14 | 33.3 | 100 | 0.769 | Very Large |
| 42 | 60 | 95.8 | 0.165 | Small | 24 | 71.4 | 66.7 | 0.009 | None |
| 43 | 48.1 | 95.8 | 0.903 | Very Large | 34 | 55.6 | 83.3 | 0.419 | Medium |
| 44 | 54.5 | 79.2 | 0.123 | Small | 44 | 54.5 | 79.2 | 0.123 | Small |
| Mean | 51.0 | 84.9 | 0.493 | | | | | | |
| Min. | 25.9 | 57.1 | 0.009 | | | | | | |
| Max. | 83.3 | 100.0 | 0.943 | | | | | | |

[a] $R^2$ was divided into five categories of effect sizes: none ($R^2$<.1), small (.1 ≤ $R^2$ <.3), medium (.3 ≤ $R^2$ <.5), large (.5 ≤ $R^2$ <.75), and very large ($R^2$ ≥.75).
[b] Percent correct for the DTs by two-year-old children
[c] Percent correct for the DTs by children who were five years or older

Table 5. Number of Correct and Incorrect Judgments for DT Productions in High Words

| | % (#) of Productions | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Type of Judgment | C2 | | C3 | | C4 | | C5+ | | Adults | |
| Correct (3) [a] | 35.3[e] | (49[f]) | 47.4 | (73) | 43.0 | (65) | 72.2 | (78) | 93.2 | (164) |
| Incorrect (3) [b] | 31.7 | (44) | 19.5 | (30) | 23.8 | (36) | 7.4 | (8) | 0.6 | (1) |
| Incorrect (2) [c] | 15.1 | (21) | 11.0 | (17) | 15.9 | (24) | 3.7 | (4) | 0.6 | (1) |
| Incorrect (1) [d] | 18.0 | (25) | 22.1 | (34) | 17.2 | (26) | 16.7 | (18) | 5.7 | (10) |
| DTs incorrectly identified by 3 judges and all 3 judges selected the same wrong DT | 15.1 | (21) | 11.0 | (17) | 11.3 | (17) | 3.7 | (4) | 0.6 | (1) |
| Total DT productions | 100 | (139) | 100 | (154) | 100 | (151) | 100 | (108) | 100 | (176) |

[a] Correct (3): DTs correctly identified by all 3 judges
[b] Incorrect (3): DTs incorrectly identified by all 3 judges
[c] Incorrect (2): DTs incorrectly identified by 2 judges
[d] Incorrect (1): DTs incorrectly identified by 1 judge
[e] Percent of total productions
[f] Number of total productions

Table 6. Judges' Responses to the DT Productions in High Words Produced by 2-year- to 4-year-old Children

| | Judges' Responses (%) (Correct responses are in black cells. Substitution patterns accounted for >10% of total judgments are highlighted) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target DT | 11 | 12 | 13 | 14 | 21 | 22 | 23 | 24 | 31 | 32 | 34 | 41 | 42 | 43 | 44 |
| 11 | **53** | 1 | 10 | 4 | 14 | | | | | 1 | | 16 | 1 | | |
| 12 | 2 | **45** | 20 | 1 | 1 | | 9 | | | 10 | | 2 | 6 | | 3 |
| 13 | 1 | 5 | **63** | 10 | | 2 | 9 | | | 1 | | 1 | 4 | 2 | 1 |
| 14 | 2 | | | **65** | | 2 | 4 | 11 | | | | 5 | 2 | | 11 |
| 21 | 4 | | | 4 | **77** | | 1 | 2 | 6 | | 4 | 1 | 1 | | |
| 22 | | 6 | 7 | 2 | 3 | **49** | 19 | 1 | | 3 | | 5 | 4 | | |
| 23 | 1 | 4 | 26 | 7 | | | **62** | | | | | | | | |
| 24 | 1 | | | 2 | 6 | 2 | 16 | **67** | | 1 | 4 | | | | |
| 31 | 3 | | | | 13 | | | 2 | **60** | 2 | 5 | 13 | 2 | | |
| 32 | | 1 | 1 | 3 | 1 | 1 | | 1 | 2 | **80** | 4 | 3 | 1 | | 1 |
| 34 | | | | | 6 | | 2 | 12 | 10 | | **57** | 4 | 1 | | 8 |
| 41 | 23 | | | 1 | 4 | 1 | 1 | 1 | 2 | 2 | | **58** | 2 | 1 | 2 |
| 42 | | 7 | 4 | | | 1 | 1 | 1 | | 8 | | 12 | **51** | 13 | 1 |
| 43 | | 2 | 13 | | | | | | | 2 | | | 16 | **63** | 3 |
| 44 | 5 | 1 | 1 | 22 | | | 1 | 9 | | | 6 | 5 | 2 | | **49** |

# Appendices

*Appendix A. Language Scores and Demographic Background of Child Participants*

| | | | | Chinese Scores (Percentile Rank) | | | English Scores (Percentile Rank) | | | Difference of | Experience (Months) in | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ID# | Age | Gender | AC[a] | EC[b] | Total | AC[a] | EC[b] | Total | Total Scores | China / Taiwan | Chinese Schools | English Schools |
| 1 | UC67 | 2;1 | M | NA | NA | NA | NA | NA | NA | NA | 0 | 0 | 0 |
| 2 | UC66 | 2;2 | F | NA | NA | NA | NA | NA | NA | NA | 25 | 0 | 0 |
| 3 | UC38 | 2;4 | F | 9 | 45 | 21 | 1 | 8 | 1 | 20 | 0 | 0 | 0 |
| 4 | UC34 | 2;5 | F | 74 | 92 | 85 | 3 | 16 | 5 | 80 | 1 | 1 | 0 |
| 5 | UC36 | 2;7 | M | 15 | 74 | 51 | 1 | 2 | 1 | 50 | 24 | 0 | 0 |
| 6 | UC68 | 2;7 | M | 53 | 48 | 47 | 5 | 4 | 3 | 44 | 3[c] | 0 | 0 |
| 7 | UC01 | 2;8 | F | 57 | 84 | 79 | 7 | 4 | 4 | 75 | 0 | 8 | 0 |
| 8 | UC35 | 2;8 | F | 43 | 56 | 51 | 16 | 4 | 6 | 45 | 0 | 0 | 10 |
| 9 | UC46 | 2;8 | M | 36 | 68 | 54 | 1 | 3 | 1 | 53 | 8 | 1 | 0 |
| 10 | UC32 | 2;9 | F | 18 | 52 | 37 | 1 | 3 | 1 | 36 | 12 | 0 | 0 |
| 11 | UC16 | 2;11 | M | 43 | 84 | 70 | 1 | 3 | 1 | 69 | 0 | 4 | 0 |
| 12 | UC44 | 2;11 | F | 53 | 73 | 64 | 1 | 4 | 1 | 63 | 0 | 0 | 0 |
| 13 | UC42 | 3;0 | F | 36 | 73 | 59 | 1 | 1 | 1 | 58 | 18 | 6 | 0 |
| 14 | UC07 | 3;3 | F | 43 | 84 | 70 | 23 | 4 | 8 | 62 | 33 | 5 | 0 |

| | | | | Chinese Scores (Percentile Rank) | | | English Scores (Percentile Rank) | | | Difference of | Experience (Months) in | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ID# | Age | Gender | AC[a] | EC[b] | Total | AC[a] | EC[b] | Total | Total Scores | China / Taiwan | Chinese Schools | English Schools |
| 15 | UC10 | 3;3 | F | 85 | 72 | 82 | 1 | 1 | 1 | 81 | 0 | 13 | 0 |
| 16 | UC47 | 3;3 | F | 85 | 92 | 92 | 1 | 1 | 1 | 91 | 3 | 0 | 1 |
| 17 | UC48 | 3;3 | M | 11 | 68 | 39 | 2 | 18 | 6 | 33 | 0 | 15 | 0 |
| 18 | UC65 | 3;4 | F | 53 | 52 | 51 | 16 | 4 | 6 | 45 | 24 | 0 | 3 |
| 19 | UC33 | 3;6 | M | 15 | 48 | 30 | 3 | 5 | 2 | 28 | 0 | 15 | 0 |
| 20 | UC62 | 3;6 | M | 90 | 93 | 93 | 45 | 5 | 16 | 77 | 0 | 0 | 0 |
| 21 | UC26 | 3;7 | M | 54 | 80 | 70 | 1 | 1 | 1 | 69 | 43 | 0 | 0 |
| 22 | UC17 | 3;8 | M | 82 | 80 | 78 | 10 | 1 | 3 | 75 | 6[c] | 0 | 12 |
| 23 | UC29 | 3;11 | F | 50 | 45 | 45 | 1 | 1 | 1 | 44 | 0 | 0 | 0 |
| 24 | UC39 | 3;11 | F | 65 | 85 | 78 | 1 | 2 | 1 | 77 | 43 | 0 | 2 |
| 25 | UC55 | 3;11 | M | 65 | 99 | 89 | 32 | 7 | 14 | 75 | 0 | 11 | 0 |
| 26 | UC41 | 4;0 | F | 54 | 74 | 67 | 14 | 9 | 9 | 58 | 2 | 11 | 0 |
| 27 | UC45 | 4;0 | F | 71 | 91 | 86 | 1 | 1 | 1 | 85 | 42 | 0 | 0 |
| 28 | UC23 | 4;1 | M | 20 | 62 | 35 | 5 | 1 | 1 | 34 | 0 | 20 | 0 |
| 29 | UC28 | 4;2 | F | 9 | 58 | 26 | 23 | 1 | 3 | 23 | 27 | 11 | 0 |
| 30 | UC72 | 4;2 | M | 40 | 89 | 69 | 4 | 1 | 1 | 68 | 30 | 0 | 14 |

| | | | | Chinese Scores (Percentile Rank) | | | English Scores (Percentile Rank) | | | Difference of | Experience (Months) in | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ID# | Age | Gender | AC[a] | EC[b] | Total | AC[a] | EC[b] | Total | Total Scores | China / Taiwan | Chinese Schools | English Schools |
| 31 | UC43 | 4;7 | M | 17 | 46 | 35 | 1 | 1 | 1 | 34 | 31 | 0 | 0 |
| 32 | UC70 | 4;7 | M | 74 | 90 | 85 | 1 | 1 | 1 | 84 | 29 | 0 | 0 |
| 33 | UC73 | 4;7 | M | 48 | 60 | 55 | 5 | 1 | 1 | 54 | 0 | 36 | 14 |
| 34 | UC75 | 4;8 | F | 17 | 62 | 30 | 14 | 2 | 5 | 25 | 5 | 0 | 12 |
| 35 | UC56 | 4;9 | F | 25 | 51 | 35 | 21 | 14 | 16 | 19 | 6 | 0 | 0 |
| 36 | UC64 | 4;10 | F | 21 | 40 | 30 | 13 | 8 | 8 | 22 | 0 | 12 | 8 |
| 37 | UC22 | 5;1 | F | 79 | 55 | 67 | 21 | 1 | 3 | 64 | 19 | 16 | 0 |
| 38 | UC54 | 5;1 | F | 68 | 81 | 76 | 7 | 1 | 1 | 75 | 3[c] | 0 | 8 |
| 39 | UC06 | 5;3 | F | 27 | 99 | 62 | 12 | 1 | 1 | 61 | 0 | 39 | 0 |
| 40 | UC08 | 5;4 | F | 29 | 71 | 51 | 16 | 1 | 1 | 50 | 57 | 5 | 0 |
| 41 | UC53 | 5;4 | F | 11 | 99 | 48 | 12 | 1 | 2 | 46 | 2 | 36 | 0 |
| 42 | UC18 | 5;5 | F | 7 | 71 | 29 | 21 | 10 | 13 | 16 | 0 | 0 | 17 |
| 43 | UC57 | 6;3 | F | 38 | 82 | 62 | 42 | 8 | 19 | 43 | 24 | 0 | 10 |
| 44 | UC52 | 6;7 | M | 94 | 82 | 92 | 1 | 1 | 1 | 91 | 78 | 0 | 0 |
| | Min. | 2;1 | M=17 | 7 | 40 | 21 | 1 | 1 | 1 | 16 | 0 | 0 | 0 |
| | Max. | 6;7 | F=27 | 94 | 99 | 93 | 45 | 18 | 19 | 91 | 78 | 39 | 17 |

| | | | Chinese Scores (Percentile Rank) | | | English Scores (Percentile Rank) | | | Difference of | Experience (Months) in | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID# | Age | Gender | AC[a] | EC[b] | Total | AC[a] | EC[b] | Total | Total Scores | China / Taiwan | Chinese Schools | English Schools |
| Mean | 3.9 | | 44.9 | 71.7 | 58.9 | 9.7 | 4.0 | 4.1 | 54.8 | 14.3 | 6.0 | 2.5 |
| SD | 1.1 | | 25.8 | 17.2 | 21.1 | 11.2 | 4.3 | 4.9 | 21.7 | 18.9 | 10.2 | 4.9 |

[a] Auditory Comprehension
[b] Expressive Communication
[c] Stayed in Taiwan

*Appendix B. Production Rates of the Target Words*

| | Tone | PinYin | Chinese | Meaning | % of Production | | Usable Child Produced Tokens[g] | Usable Adult Produced Tokens | Category[h] |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | in WFT1[a] & WFT2[a] | by Children in this Study[b] | | | |
| 1 | 11 | XiGua | 西瓜 | Watermelon | 80.8[c] (21[d] /26[e]) | 100.0[c] (44[d] /44[e]) | 39 | 11 | High |
| 2 | 12 | ShuaYa | 刷牙 | Brush teeth | 59.3 (35/59) | 97.7 (43/44) | 36 | 12 | High |
| 3 | 13 | HeShui | 喝水 | Drink water | 69.5 (41/59) | 90.9 (40/44) | 34 | 12 | High |
| 4 | 14 | JiDan | 雞蛋 | Egg | 48.5 (16/33)[f] | 68.2 (30/44) | 23 | 12 | High |
| 5 | 21 | MaoJin | 毛巾 | Towel | 45.5 (15/33) | 77.3 (34/44) | 34 | 12 | High |
| 6 | 22 | HuDie | 蝴蝶 | Butterfly | 72.9 (43/59) | 95.5 (42/44) | 40 | 12 | High |
| 7 | 23 | PingGuo | 蘋果 | Apple | 79.7 (47/59) | 93.2 (41/44) | 34 | 11 | High |
| 8 | 24 | PangXie | 螃蟹 | Crab | 53.7 (22/41) | 86.4 (38/44) | 35 | 11 | High |
| 9 | 31 | JianDao | 剪刀 | Scissors | 84.8 (50/59) | 95.5 (42/44) | 39 | 12 | High |
| 10 | 32 | CaoMei | 草莓 | Strawberry | 61.0 (36/59) | 88.6 (39/44) | 38 | 12 | High |
| 11 | 34 | YanJing | 眼鏡 | Eye glasses | 79.7 (47/59) | 100.0 (44/44) | 43 | 12 | High |
| 12 | 41 | DanGao | 蛋糕 | Cake | 62.7 (37/59) | 90.9 (40/44) | 38 | 12 | High |
| 13 | 42 | QiQiu | 氣球 | Balloon | 66.1 (39/59) | 95.5 (42/44) | 38 | 11 | High |
| 14 | 43 | DianNao | 電腦 | Computer | 67.8 (40/59) | 93.2 (41/44) | 39 | 12 | High |

| | Tone | PinYin | Chinese | Meaning | % of Production in WFT1[a] & WFT2[a] | % of Production by Children in this Study[b] | Usable Child Produced Tokens[g] | Usable Adult Produced Tokens | Category[h] |
|---|---|---|---|---|---|---|---|---|---|
| 15 | 44 | DianHua | 電話 | Telephone | 80.8 (21/26) | 95.5 (42/44) | 42 | 12 | High |
| 16 | 11 | WuGui | 烏龜 | Turtle | 72.7[c] (24[d]/33[e]) | 93.2 [c] (41[d] /44[e]) | 36 | 11 | Low |
| 17 | 12 | GongYuan | 公園 | Park | 40.0 (13/33) | 61.4 (27/44) | 23 | 12 | Low |
| 18 | 13 | QianBi | 鉛筆 | Pencil | 50.9 (30/59) | 72.7 (32/44) | 30 | 12 | Low |
| 19 | 14 | LaLian | 拉鏈 | Zipper | 36.4 (12/33) | 47.7 (21/44) | 18 | 12 | Low |
| 20 | 21 | YaGao | 牙膏 | Tooth paste | 47.5 (28/59) | 81.8 (36/44) | 29 | 12 | Low |
| 21 | 22 | ChuFang | 廚房 | Kitchen | 27.9 (16/59) | 65.9 (29/44) | 23 | 12 | Low |
| 22 | 23 | NiuNai | 牛奶 | Milk | 72.7 (24/33)[f] | 81.8 (36/44) | 34 | 12 | Low |
| 23 | 24 | WanJu | 玩具 | Toys | 57.6 (19/33)[f] | 81.8 (36/44) | 29 | 12 | Low |
| 24 | 31 | BingGan | 餅乾 | Biscuits | 66.1 (39/59) | 88.6 (39/44) | 35 | 11 | Low |
| 25 | 32 | KongLong | 恐龍 | Dinosaur | 36.4 (12/33) | 72.7 (32/44) | 27 | 12 | Low |
| 26 | 34 | ShouTao | 手套 | Gloves | 67.8 (40/59) | 88.6 (39/44) | 36 | 11 | Low |
| 27 | 41 | MianBao | 麵包 | Bread | 50.9 (30/59) | 90.9 (40/44) | 36 | 12 | Low |
| 28 | 42 | MianTiao | 麵條 | Noodles | 33.3 (11/33)[f] | 59.1 (26/44) | 22 | 12 | Low |
| 29 | 43 | BaoZhi | 報紙 | Newspaper | 66.1 (39/59) | 93.2 (41/44) | 34 | 12 | Low |
| 30 | 44 | DaXiang | 大象 | Elephant | 74.6 (44/59) | 88.6 (39/44) | 35 | 12 | Low |

[a] WFT1 and WFT2 are the two word familiarity testing pilots carried out in three preschools.
[b] The percent and number of productions were based on the number of productions of the target words in any utterance position by the 44 children in the present study.
[c] Percent of children who produced the target word
[d] Number of children who produced the target word
[e] Total number of children who were presented the picture of the target word
[f] Words that were tested with 1 picture presentation in WFT1 and 2 different picture presentations in WFT2. Thus, the production rates in WFT1 were discarded. The picture presentations with higher production rates in WFT2 are presented.
[g] Usable tokens are the number of children's productions used in this study for tone judgments. They are the target words that were produced in isolation and were not imitations or noisy, playful, clipped or soft tokens.
[h] For the two words for each DT, the one that contributed more usable tokens for this study was categorized as High Attempted Word. The other one was categorized as Low Attempted Word. There are a few exceptions to this general rule for categorizing high and low tokens (see the results section under "Comparison of Adults' and Children's Tone Production Accuracy".

*Appendix C. Categories of Mandarin Initial Consonants in the lowest to highest f0 perturbation and interruption*

| Category | Type of Initial Consonants | PinYin | IPA |
|---|---|---|---|
| 1 | Sonorants | m, n, l, | m, n, l |
| 2 | Approximants | w, r, y | w, ɹ, j |
| 3 | Unaspirated Stops | b, d, g | p, t, k |
| 4 | Unaspirated Affricates | z, zh, j | ts, tʂ, tɕ |
| 5 | Fricatives | f, s, sh, x, h | f, s, ʂ, ɕ, χ |
| 6 | Aspirated Stops | p, t, k | $p^h$, $t^h$, $k^h$ |
| 7 | Aspirated Affricates | c, ch, q | $ts^h$, $tʂ^h$, $tɕ^h$ |

*Appendix D. Types of Responses from Children*

| | Tone | PinYin | Meaning | Isolated Target Word | Target Words Not in Isolation | Non Target Words | Monosyllabic Word | Duplicated Syllables | Response in English | No Response |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11 | XiGua | Watermelon | 95.5[a] (42[b]) | 4.5 (2) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| 2 | 12 | ShuaYa | Brush teeth | 77.3 (34) | 20.5 (9) | 2.3 (1) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| 3 | 13 | HeShui | Drink water | 84.1 (37) | 6.8 (3) | 4.5 (2) | 0 (0) | 0 (0) | 0 (0) | 4.5 (2) |
| 4 | 14 | JiDan | Egg | 68.2 (30) | 0 (0) | 0 (0) | 11.4 (5) | 20.5 (9) | 0 (0) | 0 (0) |
| 5 | 21 | MaoJin | Towel | 75 (33) | 2.3 (1) | 11.4 (5) | 0 (0) | 0 (0) | 0 (0) | 11.4 (5) |
| 6 | 22 | HuDie | Butterfly | 90.9 (40) | 4.5 (2) | 2.3 (1) | 0 (0) | 0 (0) | 2.3 (1) | 0 (0) |
| 7 | 23 | PingGuo | Apple | 88.6 (39) | 4.5 (2) | 0 (0) | 0 (0) | 2.3 (1) | 2.3 (1) | 2.3 (1) |
| 8 | 24 | PangXie | Crab | 81.8 (36) | 4.5 (2) | 9.1 (4) | 0 (0) | 0 (0) | 0 (0) | 4.5 (2) |
| 9 | 31 | JianDao | Scissors | 95.5 (42) | 0 (0) | 2.3 (1) | 0 (0) | 0 (0) | 0 (0) | 2.3 (1) |
| 10 | 32 | CaoMei | Strawberry | 88.6 (39) | 0 (0) | 2.3 (1) | 0 (0) | 0 (0) | 6.8 (3) | 2.3 (1) |
| 11 | 34 | YanJing | Eye glasses | 100 (44) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) |
| 12 | 41 | DanGao | Cake | 84.1 (37) | 6.8 (3) | 2.3 (1) | 2.3 (1) | 0 (0) | 2.3 (1) | 2.3 (1) |
| 13 | 42 | QiQiu | Balloon | 93.2 (41) | 2.3 (1) | 2.3 (1) | 0 (0) | 0 (0) | 2.3 (1) | 0 (0) |
| 14 | 43 | DianNao | Computer | 90.9 (40) | 2.3 (1) | 4.5 (2) | 0 (0) | 0 (0) | 2.3 (1) | 0 (0) |
| 15 | 44 | DianHua | Telephone | 88.6 (39) | 6.8 (3) | 2.3 (1) | 0 (0) | 0 (0) | 0 (0) | 2.3 (1) |
| 16 | 11 | WuGui | Turtle | 90.9 (40) | 2.3 (1) | 2.3 (1) | 0 (0) | 0 (0) | 4.5 (2) | 0 (0) |

| | Tone | PinYin | Meaning | Isolated Target Word | Target Words Not in Isolation | Non Target Words | Monosyllabic Word | Duplicated Syllables | Response in English | No Response |
|---|---|---|---|---|---|---|---|---|---|---|
| 17 | 12 | GongYuan | Park | 50 (22) | 11.4 (5) | 18.2 (8) | 0 (0) | 0 (0) | 11.4 (5) | 9.1 (4) |
| 18 | 13 | QianBi | Pencil | 70.5 (31) | 2.3 (1) | 2.3 (1) | 22.7 (10) | 2.3 (1) | 0 (0) | 0 (0) |
| 19 | 14 | LaLian | Zipper | 45.5 (20) | 2.3 (1) | 36.4 (16) | 0 (0) | 0 (0) | 0 (0) | 15.9 (7) |
| 20 | 21 | YaGao | Tooth paste | 75 (33) | 6.8 (3) | 4.5 (2) | 0 (0) | 0 (0) | 0 (0) | 13.6 (6) |
| 21 | 22 | ChuFang | Kitchen | 61.4 (27) | 4.5 (2) | 15.9 (7) | 0 (0) | 0 (0) | 2.3 (1) | 15.9 (7) |
| 22 | 23 | NiuNai | Milk | 79.5 (35) | 2.3 (1) | 2.3 (1) | 4.5 (2) | 4.5 (2) | 0 (0) | 6.8 (3) |
| 23 | 24 | WanJu | Toys | 72.7 (32) | 9.1 (4) | 13.6 (6) | 0 (0) | 2.3 (1) | 0 (0) | 2.3 (1) |
| 24 | 31 | BingGan | Biscuits | 84.1 (37) | 4.5 (2) | 4.5 (2) | 0 (0) | 2.3 (1) | 0 (0) | 4.5 (2) |
| 25 | 32 | KongLong | Dinosaur | 70.5 (31) | 2.3 (1) | 11.4 (5) | 0 (0) | 0 (0) | 11.4 (5) | 4.5 (2) |
| 26 | 34 | ShouTao | Gloves | 86.4 (38) | 2.3 (1) | 4.5 (2) | 0 (0) | 0 (0) | 0 (0) | 6.8 (3) |
| 27 | 41 | MianBao | Bread | 88.6 (39) | 2.3 (1) | 2.3 (1) | 0 (0) | 0 (0) | 0 (0) | 6.8 (3) |
| 28 | 42 | MianTiao | Noodles | 56.8 (25) | 2.3 (1) | 2.3 (1) | 15.9 (7) | 20.5 (9) | 0 (0) | 2.3 (1) |
| 29 | 43 | BaoZhi | Newspaper | 93.2 (41) | 0 (0) | 4.5 (2) | 0 (0) | 0 (0) | 0 (0) | 2.3 (1) |
| 30 | 44 | DaXiang | Elephant | 88.6 (39) | 0 (0) | 4.5 (2) | 0 (0) | 0 (0) | 6.8 (3) | 0 (0) |
| | | | **Total** | 80.5 (1063) | 4 (53) | 5.8 (77) | 1.9 (25) | 1.8 (24) | 1.8 (24) | 4.1 (54) |

[a] Percentage of trials. Total number of trials is 1320 (44 children x 30 words).
[b] Number of trials

*Appendix E. Interjudge Correlations among the Five Judges*

E1. Pearson Rank-Order Coefficient on the Overall Scores of the 56 Adults and Children

| | J1 | J2 | J3 | J4 |
|---|---|---|---|---|
| J2 | .924** | | | |
| J3 | .863** | .887** | | |
| J4 | .930** | **.936**** | .877** | |
| J5 | .926** | **.893**** | .866** | **.903**** |

E2. Pearson Rank-Order Coefficient on the Overall Scores of the 44 Child Speakers

| | J1 | J2 | J3 | J4 |
|---|---|---|---|---|
| J2 | .876** | | | |
| J3 | .764** | .799** | | |
| J4 | .895** | **.889**** | .770** | |
| J5 | .894** | **.831**** | .785** | **.832**** |

E3. Pearson Rank-Order Coefficient on the 15 DTs Produced by the 56 Adult and Child Speakers

| | J1 | J2 | J3 | J4 |
|---|---|---|---|---|
| J2 | .014 | | | |
| J3 | .329 | .039 | | |
| J4 | -.046 | **.800**** | .261 | |
| J5 | -.082 | **.743**** | .246 | **.800**** |

E4. Pearson Rank-Order Coefficient on the 15 DTs Produced by the 44 Child Speakers

| | J1 | J2 | J3 | J4 |
|---|---|---|---|---|
| J2 | .125 | | | |
| J3 | .282 | .068 | | |
| J4 | .211 | **.800**** | .400 | |
| J5 | .286 | **.750**** | .529* | **.821**** |

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.05 level (2-tailed)
Correlations among the judges who were selected for further analysis (J2, J4, J5) are in bold.

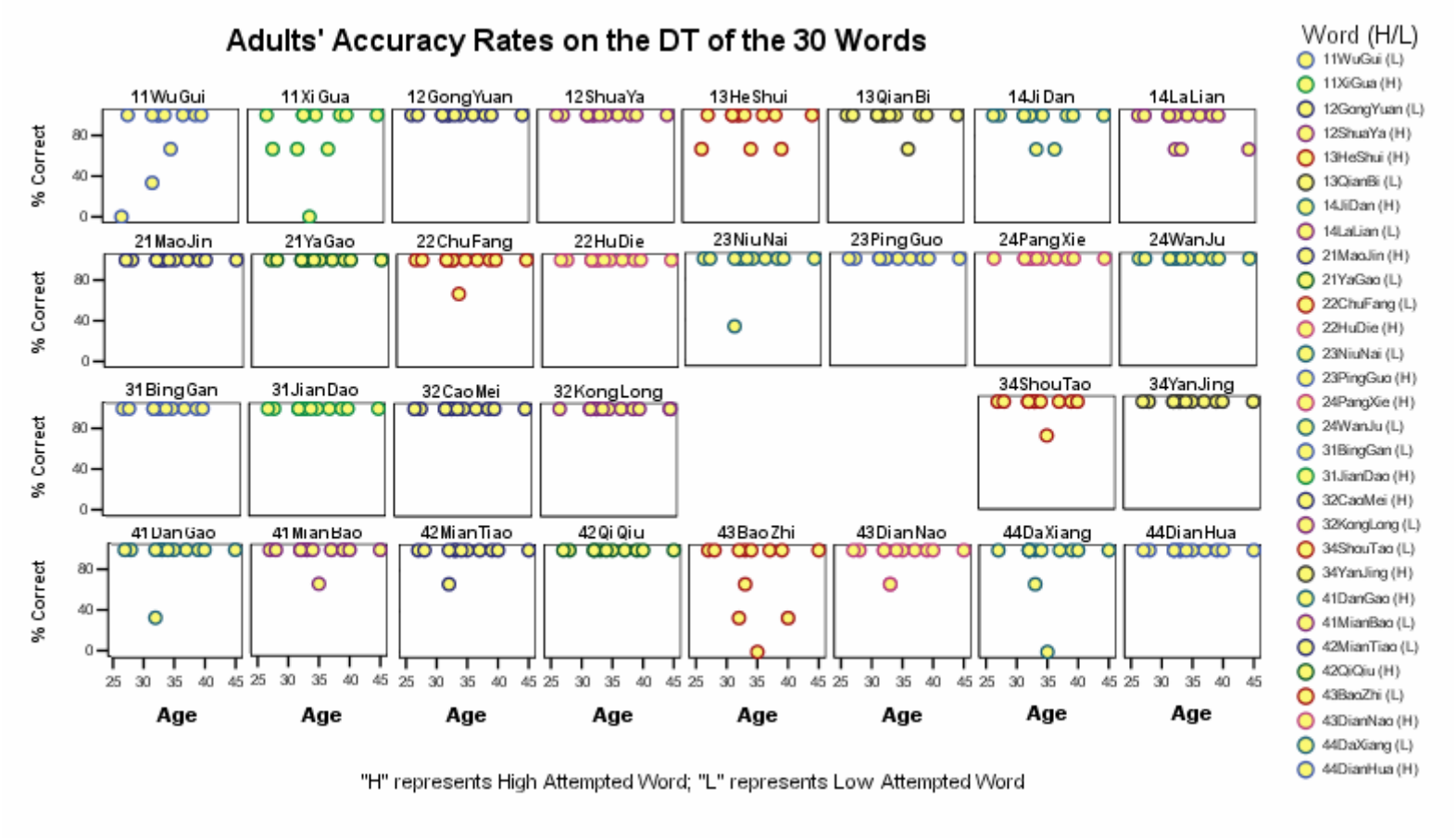
## *Appendix F. Intrajudge Correlations of the Three Selected Judges*

Pearson Rank-Order Correlation Coefficients of the Test-Retest Reliability of the Judges on their Overall Accuracy of the Speakers and the 15 Disyllabic Tones

| | Overall Scores of All Speakers | Overall Scores of Child Speakers | Accuracy of 15 DTs by All Speakers | Accuracy of 15 DTs by Child Speakers |
|---|---|---|---|---|
| J2 | .967** | .964** | .914** | .936** |
| J4 | .912** | .833* | .338 | .286 |
| J5 | .946** | .893** | .733** | .732** |

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.05 level (2-tailed)

*Appendix G. Adults' Accuracy on 30 Target Words*



*Appendix H. Number of High and Low Words Produced*

| Age Group | High Word | Low Word | High + Low Word |
|---|---|---|---|
| C2 | 139 | 95 | 234 |
| C3 | 154 | 123 | 277 |
| C4 | 151 | 124 | 275 |
| C5+ | 108 | 105 | 213 |
| All Children | 552 | 447 | 999 |
| Adults | 176 | 177 | 353 |
| Children & Adults | 728 | 624 | 1352 |

*Appendix I. Results of Mann-Whitney U Test on Children's vs. Adults' Accuracy on the 15 Disyllabic Tones*

| | All Words | | | | High Words | | | |
|---|---|---|---|---|---|---|---|---|
| DT | N | z | p | Mean (%) | N | z | p | Mean (%) |
| 11 | 53 | -1.991 | 0.046* | 62 [a] | 50 | -1.594 | 0.111 | 62 [a] |
| 12 | 50 | -4.636 | 0.000** | 47 | 48 | -4.218 | 0.000** | 47 |
| 13 | 51 | -2.233 | 0.026* | 69 | 46 | -2.206 | 0.027* | 64 |
| 14 | 45 | -1.203 | 0.229 | 74 | 35 | -1.976 | 0.092 | 71 |
| 21 | 51 | -2.922 | 0.003** | 74 | 46 | -2.082 | 0.037* | 81 |
| 22 | 54 | -4.046 | 0.000** | 45 | 52 | -3.431 | 0.001** | 56 |
| 23 | 52 | -2.650 | 0.008** | 73 | 45 | -2.370 | 0.018* | 69 |
| 24 | 48 | -3.435 | 0.001** | 64 | 46 | -2.797 | 0.005** | 67 |
| 31 | 55 | -4.085 | 0.000** | 65 | 51 | -3.281 | 0.001** | 65 |
| 32 | 52 | -2.229 | 0.026* | 83 | 50 | -2.072 | 0.038* | 81 |
| 34 | 56 | -3.569 | 0.000** | 64 | 55 | -3.368 | 0.001** | 62 |
| 41 | 53 | -2.917 | 0.004** | 69 | 50 | -2.502 | 0.012* | 63 |
| 42 | 51 | -3.520 | 0.000** | 59 | 49 | -3.199 | 0.001** | 61 |
| 43 | 54 | -2.203 | 0.028* | 63 | 51 | -2.052 | 0.040* | 70 |
| 44 | 55 | -3.041 | 0.002** | 58 | 54 | -3.435 | 0.001** | 55 |

** Significant at the 0.01 level (2-tailed)
* Significant at the 0.05 level (2-tailed)
[a] Mean percent correct of 2- to 6-year-old children's productions

*Appendix J. Number of Children in Each Age Group whose Accuracy Rates of the 15 DTs in High Words were Adult-like*

Table J1. All 30 Words Included

| | | 11 C | 12 NC | 13 NC | 14 C | 21 C | 22 NC | 23 NC | 24 C | 31 NC | 32 C | 34 NC | 41 NC | 42 C | 43 C | 44 NC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C2 | # within 95%CI [a] | 3 | 0 | 3 | 2 | 1 | 3 | 4 | 4 | 2 | 9 | 2 | 3 | 4 | 1 | 2 |
| | # produced the DT [b] | 9 | 9 | 9 | 8 | 8 | 12 | 9 | 8 | 12 | 12 | 12 | 12 | 11 | 10 | 11 |
| | % within 95%CI [c] | 33 | 0 | 33 | 25 | 13 | 25 | 44 | 50 | 17 | 75 | 17 | 25 | 36 | 10 | 18 |
| C3 | # within 95%CI | 7 | 2 | 5 | 6 | 7 | 2 | 5 | 4 | 6 | 5 | 5 | 5 | 4 | 7 | 5 |
| | # produced the DT | 13 | 12 | 12 | 9 | 12 | 11 | 12 | 9 | 13 | 10 | 13 | 10 | 10 | 13 | 13 |
| | % within 95%CI | 54 | 17 | 42 | 67 | 58 | 18 | 42 | 44 | 46 | 50 | 39 | 50 | 40 | 54 | 39 |
| C4 | # within 95%CI | 7 | 0 | 4 | 5 | 6 | 2 | 5 | 3 | 3 | 7 | 2 | 5 | 0 | 4 | 3 |
| | # produced the DT | 11 | 10 | 10 | 9 | 11 | 11 | 11 | 11 | 11 | 10 | 11 | 11 | 10 | 11 | 11 |
| | % within 95%CI | 64 | 0 | 40 | 56 | 55 | 18 | 46 | 27 | 27 | 70 | 18 | 46 | 0 | 36 | 27 |
| C5 | # within 95%CI | 8 | 3 | 5 | 5 | 6 | 3 | 5 | 3 | 0 | 6 | 5 | 3 | 5 | 5 | 5 |
| | # produced the DT | 8 | 7 | 8 | 7 | 8 | 8 | 8 | 8 | 7 | 8 | 8 | 8 | 8 | 8 | 8 |
| | % within 95%CI | 100 | 43 | 63 | 71 | 75 | 38 | 63 | 38 | 0 | 75 | 63 | 38 | 63 | 63 | 63 |
| A | Adult 95%CI | 66.4 | 95.6 [d] | 89.2 | 87.6 | 95.6 [d] | 95.6 | 91.1 | 95.6 [d] | 95.6 [d] | 95.6 [d] | 85.6 | 86.7 | 95.6 | 74.3 | 85.0 |

a Number of children in the age group whose accuracy rate for the DT was equal to or higher than the lower bound of the 95% confident interval of adults' scores.
b Total number of children who produced one or two words for the DT and whose productions were included in this study
c Percent of children whose DT accuracy rates were higher than or equal to the lower bound of the 95% confidence interval of the adults' scores.

[d] All adults' productions were judged with 100% accuracy. Thus the lower bound of the 95% confidence interval of the adults' scores was set at .95.6%.

Table J2. High Words Only

| | | 11 C | 12 NC | 13 NC | 14 C | 21 C | 22 NC | 23 NC | 24 C | 31 NC | 32 C | 34 NC | 41 NC | 42 C | 43 C | 44 NC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C2 | # within 95%CI[a] | 2 | 0 | 3 | 0 | 2 | 4 | 2 | 5 | 6 | 9 | 3 | 4 | 4 | 2 | 4 |
| | # produced the DT[b] | 9 | 9 | 7 | 6 | 7 | 12 | 6 | 7 | 11 | 12 | 12 | 11 | 10 | 9 | 11 |
| | % within 95%CI[c] | 22 | 0 | 43 | 0 | 29 | 33 | 33 | 71 | 55 | 75 | 25 | 36 | 40 | 22 | 36 |
| C3 | # within 95%CI | 6 | 3 | 3 | 5 | 6 | 3 | 7 | 4 | 4 | 6 | 5 | 6 | 4 | 7 | 6 |
| | # produced the DT | 11 | 10 | 10 | 6 | 9 | 11 | 11 | 9 | 11 | 10 | 12 | 9 | 10 | 12 | 13 |
| | % within 95%CI | 55 | 30 | 30 | 83 | 67 | 27 | 64 | 44 | 36 | 60 | 42 | 67 | 40 | 58 | 46 |
| C4 | # within 95%CI | 7 | 2 | 5 | 3 | 9 | 4 | 6 | 6 | 3 | 7 | 4 | 4 | 1 | 5 | 2 |
| | # produced the DT | 11 | 10 | 10 | 7 | 11 | 10 | 10 | 11 | 10 | 10 | 11 | 10 | 10 | 10 | 10 |
| | % within 95%CI | 64 | 20 | 50 | 43 | 82 | 40 | 60 | 55 | 30 | 70 | 36 | 40 | 10 | 50 | 20 |
| C5 | # within 95%CI | 8 | 3 | 4 | 4 | 7 | 5 | 6 | 3 | 4 | 5 | 6 | 5 | 7 | 7 | 5 |
| | # produced the DT | 8 | 7 | 7 | 4 | 7 | 7 | 7 | 8 | 7 | 6 | 8 | 8 | 8 | 8 | 8 |
| | % within 95%CI | 100 | 43 | 57 | 100 | 100 | 71 | 86 | 38 | 57 | 83 | 75 | 63 | 88 | 88 | 63 |
| A | Adult 95%CI | 60.9 | 95.6[d] | 82.1 | 86.2 | 95.6[d] | 95.6[d] | 95.6[d] | 95.6[d] | 95.6[d] | 95.6[d] | 95.6[d] | 82.2 | 95.6[d] | 86.2 | 95.6[d] |

[a] Number of children in the age group whose accuracy rate for the DT was equal to or higher than the lower bound of the 95% confident interval of adults' scores.
[b] Total number of children who produced the High Word for the DT and whose productions were included in this study
[c] Percent of children whose DT accuracy rates were higher than or equal to the lower bound of the 95% confidence interval of the adults' scores.
[d] All adults' productions were judged with 100% accuracy. Thus the lower bound of the 95% confidence interval of the adults' scores was set at .95.6%.

*Appendix K. Number of Productions and Number of Same Speakers for High and Low Words*

| | # of Productions in C2 | | | # of Productions in C3 | | | # of Productions in C4 | | | # of Productions in C5 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DT | High Word | Low Word | # of Same Speaker | High Word | Low Word | # of Same Speaker | High Word | Low Word | # of Same Speaker | High Word | Low Word | # of Same Speaker |
| 11 | 9 | 8 | 8 | 11 | 10 | 8 | 11 | 11 | 11 | 8 | 7 | 7 |
| 12 | 9 | 4 | 4 | 10 | 7 | 5 | 10 | 6 | 6 | 7 | 6 | 6 |
| 13 | 7 | 7 | 5 | 10 | 8 | 6 | 10 | 8 | 8 | 7 | 7 | 6 |
| 14 | 6 | 4 | 2 | 6 | 3 | 0 | 7 | 5 | 3 | 4 | 6 | 3 |
| 21 | 7 | 4 | 3 | 9 | 9 | 6 | 11 | 9 | 9 | 7 | 7 | 6 |
| 22 | 12 | 5 | 5 | 11 | 6 | 6 | 10 | 5 | 4 | 7 | 7 | 6 |
| 23 | 6 | 5 | 2 | 11 | 11 | 10 | 10 | 10 | 9 | 7 | 8 | 7 |
| 24 | 7 | 3 | 2 | 9 | 8 | 8 | 11 | 11 | 11 | 8 | 7 | 7 |
| 31 | 11 | 9 | 8 | 11 | 11 | 9 | 10 | 8 | 7 | 7 | 7 | 7 |
| 32 | 12 | 6 | 6 | 10 | 8 | 8 | 10 | 7 | 7 | 6 | 6 | 4 |
| 34 | 12 | 9 | 9 | 12 | 10 | 9 | 11 | 9 | 9 | 8 | 8 | 8 |
| 41 | 11 | 8 | 7 | 9 | 10 | 9 | 10 | 10 | 9 | 8 | 8 | 8 |
| 42 | 10 | 3 | 2 | 10 | 6 | 6 | 10 | 7 | 7 | 8 | 6 | 5 |
| 43 | 9 | 10 | 9 | 12 | 8 | 7 | 10 | 8 | 7 | 8 | 8 | 8 |
| 44 | 11 | 10 | 10 | 13 | 8 | 8 | 10 | 10 | 9 | 8 | 7 | 7 |

*Appendix L. Judges' Responses to Adults' DT Productions in High Words*

| | Judges' Responses (%) (Correct responses are in black cells.) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target DT | 11 | 12 | 13 | 14 | 21 | 22 | 23 | 24 | 31 | 32 | 34 | 41 | 42 | 43 | 44 |
| 11 | **82** | | 9 | | 3 | | | | | | | 6 | | | |
| 12 | | **100** | | | | | | | | | | | | | |
| 13 | | 8 | **92** | | | | | | | | | | | | |
| 14 | | | 3 | **94** | | | | 3 | | | | | | | |
| 21 | | | | | **100** | | | | | | | | | | |
| 22 | | | | | | **100** | | | | | | | | | |
| 23 | | | | | | | **100** | | | | | | | | |
| 24 | | | | | | | | **100** | | | | | | | |
| 31 | | | | | | | | | **100** | | | | | | |
| 32 | | | | | | | | | | **100** | | | | | |
| 34 | | | | | | | | | | | **100** | | | | |
| 41 | | 3 | | | | | | | 3 | | | **94** | | | |
| 42 | | | | | | | | | | | | | **100** | | |
| 43 | | | | | | | 3 | | | | | | 3 | **94** | |
| 44 | | | | | | | | | | | | | | | **100** |

*Appendix M. Judges' Responses to All Children's DT Productions in High Words*

| | Judges' Responses (%)<br>(Correct responses are in black cells. Substitution patterns accounting for >10% of total judgments are highlighted) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target DT | 11 | 12 | 13 | 14 | 21 | 22 | 23 | 24 | 31 | 32 | 34 | 41 | 42 | 43 | 44 |
| 11 | **62** | 1 | 8 | 4 | 11 | | | | | 1 | | 13 | 1 | | |
| 12 | 2 | **47** | 19 | 1 | 1 | 2 | 8 | | | 11 | | 2 | 5 | | 3 |
| 13 | 2 | 7 | **64** | 8 | 1 | 2 | 9 | | | 1 | | 1 | 3 | 2 | 1 |
| 14 | 1 | | | **71** | | 1 | 3 | 9 | | | | 4 | 1 | | 9 |
| 21 | 3 | | | 3 | **81** | | 1 | 2 | 5 | | 3 | 1 | 1 | | |
| 22 | | 5 | 6 | 2 | 3 | **56** | 18 | 1 | | 3 | | 4 | 3 | | |
| 23 | 1 | 3 | 22 | 6 | | | **69** | | | | | | | | |
| 24 | 1 | | | 2 | 6 | 2 | 16 | **67** | | 1 | 6 | | | | |
| 31 | 3 | | | | 10 | | | 2 | **65** | 2 | 4 | 13 | 2 | | |
| 32 | 1 | 1 | 1 | 3 | 1 | 1 | | 1 | 3 | **81** | 4 | 3 | 1 | | 2 |
| 34 | | | | | 5 | | 2 | 10 | 9 | | **62** | 4 | 1 | | 9 |
| 41 | 19 | | | 1 | 4 | 1 | 1 | 1 | 4 | 2 | | **63** | 2 | 1 | 2 |
| 42 | | 6 | 4 | | | 1 | 1 | 1 | | 6 | | 10 | **61** | 11 | 1 |
| 43 | | 2 | 10 | | | | | | | 2 | | | 14 | **70** | 3 |
| 44 | 4 | 1 | 1 | 21 | | | 1 | 7 | | | 5 | 4 | 2 | | **55** |

*Appendix N. Judges' Responses to 2-year-old Children's DT Productions in High Words*

| | Judges' Responses (%) (Correct responses are in black cells. Substitution patterns accounting for >10% of total judgments are highlighted) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target DT | 11 | 12 | 13 | 14 | 21 | 22 | 23 | 24 | 31 | 32 | 34 | 41 | 42 | 43 | 44 |
| 11 | **33** | | 11 | 11 | 30 | | | | | | | 11 | 4 | | |
| 12 | 4 | **26** | 19 | 4 | 4 | | 15 | | | 15 | | 4 | | | 11 |
| 13 | | | **52** | 29 | | | 10 | | | | | 5 | 5 | | |
| 14 | 6 | | | **33** | | 6 | 11 | 17 | | | | 17 | 6 | | 6 |
| 21 | 10 | | | | **52** | | 5 | 10 | 5 | | 10 | 5 | 5 | | |
| 22 | | 3 | 14 | 6 | 8 | **47** | 19 | 3 | | | | | | | |
| 23 | | | 39 | 22 | | | **39** | | | | | | | | |
| 24 | | | | | | | 24 | **71** | | 5 | | | | | |
| 31 | | | | | 9 | | | 3 | **67** | 3 | 15 | | 3 | | |
| 32 | | | 3 | | | | | 3 | | **83** | 8 | | | | 3 |
| 34 | | | | | 3 | | 6 | 19 | 8 | | **56** | 6 | | | 3 |
| 41 | 39 | | | | 3 | | | 3 | | 6 | | **42** | 6 | | |
| 42 | | 3 | 7 | | | | | 3 | | 10 | | 3 | **60** | 10 | 3 |
| 43 | | | 19 | | | | | | | | | | 22 | **48** | 11 |
| 44 | 9 | 3 | 3 | 12 | | | 3 | 9 | | | 6 | | | | **55** |

*Appendix O. Judges' Responses to 3-year-old Children's DT Productions in High Words*

| | Judges' Responses (%) (Correct responses are in black cells. Substitution patterns accounting for >10% of total judgments are highlighted) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target DT | 11 | 12 | 13 | 14 | 21 | 22 | 23 | 24 | 31 | 32 | 34 | 41 | 42 | 43 | 44 |
| 11 | **58** | 3 | 18 | 3 | 9 | | | | | | | 9 | | | |
| 12 | | **63** | 17 | | | | 13 | | | | | | 7 | | |
| 13 | 3 | 13 | **67** | 3 | | | | | | 3 | | | 3 | 7 | |
| 14 | | | | **83** | | | | | | | | | | | 17 |
| 21 | | | | | **81** | | | | 15 | | 4 | | | | |
| 22 | | 3 | 6 | | | **45** | 33 | | | 3 | | 6 | 3 | | |
| 23 | 3 | | 21 | 6 | | | **70** | | | | | | | | |
| 24 | | | | | 4 | 7 | 22 | **59** | | | 7 | | | | |
| 31 | 6 | | | | 12 | | | 3 | **67** | 3 | | 9 | | | |
| 32 | | 3 | | | | 3 | | | | **80** | | 10 | 3 | | |
| 34 | | | | | 8 | | | 3 | 11 | | **64** | | 3 | | 11 |
| 41 | 11 | | | | 4 | | | | 7 | | | **78** | | | |
| 42 | | 3 | | | | | 3 | | | 7 | | 3 | **60** | 23 | |
| 43 | | 6 | 8 | | | | | | | | | | 17 | **69** | |
| 44 | 3 | | | 33 | | | | | | | 8 | | | | **56** |

*Appendix P. Judges' Responses to Four-year-old Children's DT Productions in High Words*

| | Judges' Responses (%) (Correct responses are in black cells. Substitution patterns accounting for >10% of total judgments are highlighted) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target DT | 11 | 12 | 13 | 14 | 21 | 22 | 23 | 24 | 31 | 32 | 34 | 41 | 42 | 43 | 44 |
| 11 | **64** | | | | 6 | | | | | 3 | | 27 | | | |
| 12 | 3 | **43** | 23 | | | | | | | 17 | | 3 | 10 | | |
| 13 | | | **67** | 3 | | 7 | 17 | | | | | | 3 | | 3 |
| 14 | | | | **76** | | | | 14 | | | | | | | 10 |
| 21 | 3 | | | 9 | **88** | | | | | | | | | | |
| 22 | | 13 | | | | **57** | 3 | | | 7 | | 10 | 10 | | |
| 23 | | 10 | 23 | | | | **67** | | | | | | | | |
| 24 | 3 | | | 6 | 12 | | 6 | **70** | | | 3 | | | | |
| 31 | 3 | | | | 17 | | | | **47** | | | 30 | 3 | | |
| 32 | | | | 10 | 3 | | | | 7 | **77** | 3 | | | | |
| 34 | | | | | 6 | | | 15 | 12 | | **52** | 6 | | | 9 |
| 41 | 17 | | | 3 | 7 | 3 | 3 | | | | | **57** | | 3 | 7 |
| 42 | | 13 | 7 | | | 3 | | | | 7 | | 30 | **33** | 7 | |
| 43 | | | 13 | | | | | | | 7 | | | 10 | **70** | |
| 44 | 3 | | | 17 | | | | 20 | | | 3 | 17 | 7 | | **33** |

*Appendix Q. Judges' Responses to the DT Productions in High Words of Five Years and Older Children*

| | Judges' Responses (%) (Correct responses are in black cells. Substitution patterns accounting for >10% of total judgments are highlighted) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target DT | 11 | 12 | 13 | 14 | 21 | 22 | 23 | 24 | 31 | 32 | 34 | 41 | 42 | 43 | 44 |
| 11 | **96** | | | 4 | | | | | | | | | | | |
| 12 | | **57** | 14 | | | 10 | 5 | | | 14 | | | | | |
| 13 | 5 | 14 | **67** | | 5 | | 10 | | | | | | | | |
| 14 | | | | **100** | | | | | | | | | | | |
| 21 | | | | | **100** | | | | | | | | | | |
| 22 | | | | | | **86** | 14 | | | | | | | | |
| 23 | | | 5 | | | | **95** | | | | | | | | |
| 24 | | | | | 4 | | 17 | **67** | | | 13 | | | | |
| 31 | | | | | | | | | **86** | | | 14 | | | |
| 32 | 6 | | | | | | | | 6 | **83** | | | | | 6 |
| 34 | | | | | | | | | | | **83** | 4 | | | 13 |
| 41 | 4 | | | | | | | | 13 | | | **83** | | | |
| 42 | | 4 | | | | | | | | | | | **96** | | |
| 43 | | | | | | | | | | | | | 4 | **96** | |
| 44 | | | | 21 | | | | | | | | | | | **79** |

*Appendix R. Major Substitution Patterns for DT Productions of Two- to Four-year-old Children*

| Target DT | Compatibility | Substition Pattern | Rate of Substitution | Tone Substitution in Syllable 1 | Tone Substitution in Syllable 2 |
|---|---|---|---|---|---|
| T11 | C | 11-->21 | 14% | T1-->T2 | |
| | | 11-->41 | 16% | T1-->T4 | |
| T12 | NC | 12-->13 | 20% | | T2-->T3 |
| T14 | C | 14-->24 | 11% | T1-->T2[a] | |
| | | 14-->44 | 11% | T1-->T4 | |
| T22 | NC | 22-->23 | 19% | | T2-->T3 |
| T23 | NC | 23-->13 | 26% | T2-->T1[a] | |
| T24 | C | 24-->23 | 16% | | T4-->T3 |
| T31 | NC | 31-->21 | 13% | T3-->T2[a] | |
| | | 31-->41 | 13% | T3-->T4 | |
| T34 | NC | 34-->24 | 12% | T3-->T2[a] | |
| T41 | NC | 41-->11 | 23% | T4-->T1[a] | |
| T42 | C | 42-->41 | 12% | | T2-->T1 |
| | | 42-->43 | 13% | | T2-->T3 |
| T43 | C | 43-->13 | 13% | T4-->T1[b] | |
| | | 43-->42 | 16% | | T3-->T2 |
| T44 | NC | 44-->14 | 22% | T4-->T1[a] | |

[a] The f0 level at the end of the incorrect tone in S1 was closer to the f0 level for the target tone in S2 than the target tone.

*Appendix S. Correlations of Children's Demographic Backgrounds and DT Accuracy in High Words*

| Intercorrelations, Means and Standard Deviations for the Demographic Variables | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | N | $r_s$ | p-value | $R^2$ | Range | Mean | SD |
| # of months in Chinese Schools | 19 | 0.503 | 0.028*[a] | 0.25 | 1-39 | 14.0 | 11.5 |
| # of months in English Schools | 12 | 0.184 | 0.568 | 0.077 | 1-17 | 9.3 | 5.1 |
| Chinese Receptive Percentile Rank | 42 | 0.14 | 0.375 | 0.025 | 7-94 | 44.9 | 25.8 |
| Chinese Expressive Percentile Rank | 42 | 0.168 | 0.289 | 0.03 | 40-99 | 71.7 | 17.2 |
| Chinese Total Percentile Rank | 42 | 0.127 | 0.423 | 0.014 | 21-93 | 58.9 | 21.1 |
| English Receptive Percentile Rank | 42 | 0.267 | 0.088 | 0.064 | 1-45 | 9.7 | 11.2 |
| English Expressive Percentile Rank | 42 | -0.289 | 0.063 | 0.063 | 1-18 | 4.0 | 4.3 |
| English Total Scores Percentile Rank | 42 | -0.015 | 0.926 | 0.0009 | 1-19 | 4.1 | 4.9 |
| # of months in Native Country | 27 | 0.123 | 0.541 | 0.007 | 1-78 | 22.1 | 19.1 |
| # of High and Low Words Attempted | 44 | 0.139 | 0.37 | 0.005 | 30-90 | 68.1 | 15.1 |

[a] Number of months in Chinese schools was found to be correlated with age ($R^2 = .446$, $r_s = .676$, $p = .001$).

*Appendix T. Correlations of DT Accuracy in High Words and Chinese and English School Education*

Figure T1. Chinese School Education and DT Accuracy

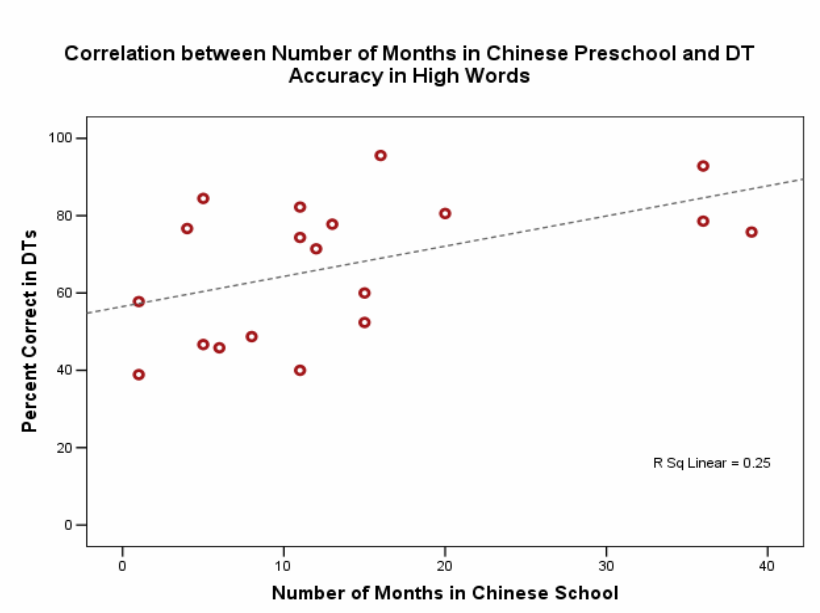
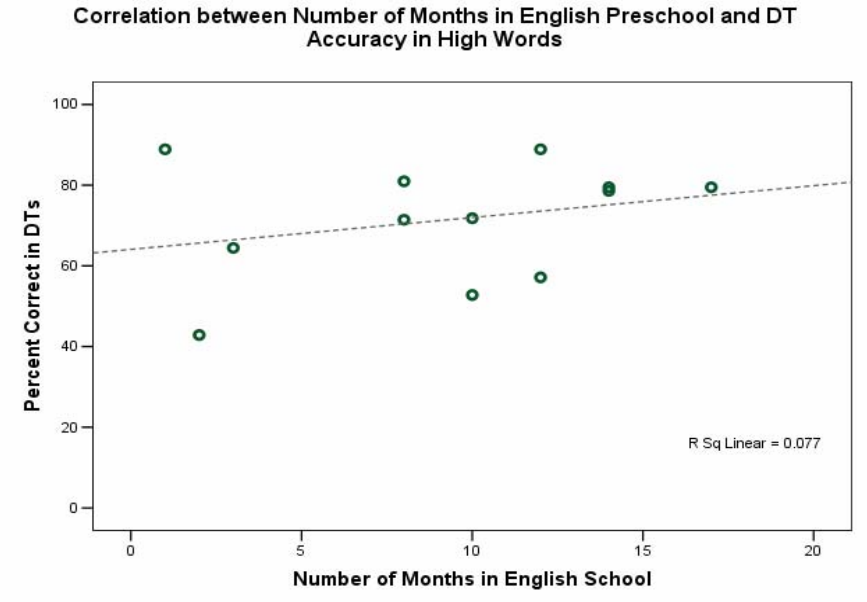



Figure T2. English School Education and DT Accuracy



# References

Adobe Systems Incorporated. (2006). *Adobe audition 2.0*. San Jose, CA:

American Speech-Language-Hearing Association. (1997). *Guidelines for audiologic screening*. MD:Author: Rockville.

Boersma, P., & Weenink, D. (1992). *Praat version 4.1.6*. University of Amsterdam, Netherlands:

Bosma, J. F. (1975). Anatomic and physiologic development of the speech apparatus. In D. B. tower (Ed.), *The nervous system: Human communication and its disorders (vol.3)* (pp. 469-481). New York: Raven Press.

Callan, D. E., Kent, R. D., Guenther, F. H., & Vorperian, H. K. (2000). An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language, and Hearing Research, 43*(3), 721-736.

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

Chao, Y. R. (1973/1951). The cantian idiolect: An analysis of the Chinese spoken by a twenty-eight-month-old child. In C. A. Ferguson, & D. I. Slobin (Eds.), *Studies of child language development* (pp. 13-33). New York: Holt, Rinehart & Winston.

Chen, Y., & Xu, Y. (2006). Production of weak elements in speech -- evidence from F(0 patterns of neutral tone in standard Chinese. *Phonetica, 63*(1), 47-75.

Clumeck, H. (1977). Topics in the acquisition of Mandarin phonology: A case study. *Papers and Reports on Child Language Development, 14*(December), 37-73.

Clumeck, H. (1980). The acquisition of tone. In G. H. Yeni-Komshian, J. F. Kavanaugh & C. A. Ferguson (Eds.), *Child phonology: Vol. 1. production* (pp. 257-275). New York: Academic Press.

Clumeck, H. V. (1977). *Studies in the acquisition of Mandarin phonology.* Unpublished doctoral dissertation, University of California, Berkerley.,

Cooper, R. P., & Aslin, R. N. (1994). Developmental differences in infant attention to the spectral properties of infant-directed speech. *Child Development, 65*(6), 1663-1677.

Crelin, E. S. (1987). *The human vocal tract: Anatomy, function, development, and evolution*. New York: Vantage Press.

Eguchi, S., & Hirsch, I. J. (1969). Development of speech sounds in children. *Acta Otolaryngologica, Suppl. 257*, 1-51.

Friederici, A. D., & Wessels, J. M. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & Psychophysics, 54*(3), 287-295.

Fu, Q. J., & Zeng, F. G. (2000). Identification of temporal envelope cues in Chinese tone recognition. *Asia Pacific Journal of Speech, Language and Hearing, 5*, 45-57.

Gårding, E., Kratochvil, P., Svantesson, J., & Zhang, J. (1986). Tone 4 and tone 3 discrimination in modern standard Chinese. *Language & Speech, 29*(3), 281-293.

Goffman, L., & Smith, A. (1999). Development and phonetic differentiation of speech movement patterns. *Journal of Experimental Psychology: Human Perception and Performance, 25*(3), 649-660.

Goffman, L., Gerken, L., & Lucchesi, J. (2007). Relations between segmental and motor variability in prosodically complex nonword sequences. *Journal of Speech, Language, and Hearing Research, 50*(2), 444-458.

Goffman, L., Smith, A., Heisler, L., & Ho, M. (2008). The breadth of coarticulatory units in children and adults. *Journal of Speech, Language, and Hearing Research,*

Goodell, E. W., & Studdert-Kennedy, M. (1993). Acoustic evidence for the development of gestural coordination in the speech of 2-year-olds: A longitudinal study. *Journal of Speech and Hearing Research, 36*(4), 707-727.

Green, J. R., Moore, C. A., Higashikawa, M., & Steeve, R. W. (2000). The physiologic development of speech motor control: Lip and jaw coordination. *Journal of Speech, Language, and Hearing Research, 43*(1), 239-255.

Hirano, M., Kurita, S., & Nakashima, T. (1981). The structure of the vocal folds. In K. N. Stevens, & M. Hirano (Eds.), *Vocal fold physiology* (1st ed., pp. 33-41). Tokyo, Japan: University of Tokyo Press.

Hirano, M., Kurita, S., & Nakashima, T. (1983). Growth, development, and aging of human vocal folds. (pp. 22-43). San Diego: College-Hill Press.

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

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

Hua, Z. (2002). *Phonological development in specific context: Studies of Chinese-speaking children*. Clevedon, England: Multilingual Matters Limited.

Hua, Z., & Dodd, B. (2000). The phonological acquisition of putonghua (modern standard Chinese). *Journal of Child Language, 27*(1), 3-42.

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development, 64*(3), 675-687.

Kahane, J. C. (1978). A morphological study of the human prepubertal and pubertal larynx. *The American Journal of Anatomy, 151*(1), 11-19.

Kahane, J. C. (1982). Growth of the human prepubertal and pubertal larynx. *Journal of Speech and Hearing Research, 25*(3), 446-455.

Kent, R. D. (1976). Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of Speech and Hearing Research, 19*(3), 421-447.

Kent, R. D. (1984). Psychobiology of speech development: Coemergence of language and a movement system. *American Journal of Physiology, 246*(6), 888-894.

Kent, R. D. (1992). The biology of phonological development. In C. A. Ferguson, L. Menn & C. Stoel-Gammon (Eds.), *Phonological development: Models, research, implications* (pp. 65-90). Timonium, MD: York Press.

Kent, R. D. (2000). Research on speech motor control and its disorders: A review and prospective. *Journal of Communication Disorders, 33*(5), 391-428.

Kent, R. D. (2004). The uniqueness of speech among motor systems. *Clinical Linguistics & Phonetics, 18*(6-8), 495-505.

Kent, R. D., & Forner, L. L. (1980). Speech segment duration in sentence recitations by children and adults. *Journal of Phonetics, 8*(2), 157-168.

Kent, R. D., & Vorperian, H. K. (1995). Development of the craniofacial-oral-laryngeal anatomy: A review. *Journal of Medical Speech-Language Pathology, 3*, 145–190.

Kent, R. D., & Vorperian, H. K. (2007). In the mouths of babes: Anatomic, motor, and sensory foundations of speech development in children. *Language disorders from a*

*developmental perspective: Essays in honor of robin S. chapman* (pp. 55-91). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kleinow, J., & Smith, A. (2006). Potential interactions among linguistic, autonomic, and motor factors in speech *Developmental Psychobiology, 48*(4), 275-287.

Lecours, A.R. (1975). Myelogenecic correlates of the development of speech and language. In E.H. Lenneberg & E. Lenneberg (Eds.), *Foundations of language development* (Vol. I, pp. 121-135). New York: Academic Press.

Li, C. N., & Thompson, S. A. (1977). The acquisition of tone in Mandarin-speaking children. *Journal of Child Language, 4*(2), 185-199.

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. Berkeley and Los Angeles, California: University of California Press.

Lin, B., & Lin, N. (1994). *Language disorder scale of preschoolers (學前兒童語言障礙評量表)*. Taipei: National Taiwan Normal University, Department of Special Education.

Locke, J. L. (1983). *Phonological acquisition and change*. New York: Academic press.

Locke, J. L. (1986). Speech perception and the emergent lexicon: An ethological approach. In P. Fletch, & M. Garman (Eds.), *Language acquisition: Studies in first*

*language development* (2nd ed., pp. 240–250). Cambridge, MA: Cambridge: Cambridge University Press.

Luo, X., & Fu, Q. J. (2004). Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants. *Journal of the Acoustical Society of America, 116*(6), 3659-3667.

Massaro, D. W., Cohen, M. M., & Tseng, C. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics, 13*, 267-290.

Nittrouer, S. (1993). The emergence of mature gestural patterns is not uniform: Evidence from an acoustic study. *Journal of Speech & Hearing Research, 36*(5), 959-972.

Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *Journal of the Acoustical Society of America, 97*(1), 520-530.

Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research, 32*(1), 120-132.

Oller, D. K., & Eilers, R. E. (1975). Phonetic expectation and transcription validity. *Phonetica, 31*(3-4), 288-304.

Ostry, D. J., Feltham, R. F., & Munhall, K. G. (1984). Characteristics of speech motor development in children. *Developmental Psychology, 20*(5), 859-871.

Seikel, J. A., King, D. W., & Drumright, D. G. (1997). *Anatomy and physiology for speech, language, and hearing* (expanded ed.). San Diego, California: Singular Publishing Group, Inc.

Shen, X. S. (1990). Tonal coarticulation in Mandarin. *Journal of Phonetics, 18*(3)

Shen, X. S. (1992). Mandarin neutral tone revisited. *Acta Linguist. Hafniensia, 24*, 131-151.

Smith, A., & Zelaznik, H. N. (2004). Development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental Psychobiology, 45*(1), 22-33.

Smith, A. (2006). Speech motor development: Integrating muscles, movements, and linguistic units. *Journal of Communication Disorders, 39*(5), 331-349.

Smith, B. L. (1978). Temporal aspects of English speech production: A developmental perspective. *Journal of Phonetics, 6*(1), 37-67.

Smith, B. L. (1991). Relationships between duration and temporal variability in children's speech. *Journal of the Acoustical Society of America, 91*, 2165-2174.

Smith, B. (2006). Phonological development in lexically precocious 2-year-olds. *Applied Psycholinguistics, 27*(3), 355.

Sony Media Software Inc. (2005). *Sound forge*, v. 8.0 [computer software]

Stathopoulos, E. T. (1995). Variability revisited: An acoustic, aerodynamic, and respiratory kinematic comparison of children and adults during speech. *Journal of Phonetics, 23*, 67-80.

Tagliaferri, B. (2005). *Paradigm* (Version SP1) [Computer software]. Perception Research Systems Inc. Available from http://www.perceptionresearchsystems.com.

Tingley, B. M., & Allen, G. D. (1975). Development of speech timing control in children. *Child Development, 46*(1), 186-194.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America, 85*(4), 1699-1707.

Titze, I. R. (1994). *Principles of voice production* (1st ed.). Englewood Cliffs, New Jersey: Prentice-Hall.

Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research, 50*(6), 1510-1545.

Walker, J. F., Archibald, L. M., Cherniak, S. R., & Fish, V. G. (1992). Articulation rate in 3- and 5-year-old children. *Journal of Speech and Hearing Research, 35*(1), 4-13.

Walsh, B., & Smith, A. (2002). Articulatory movements in adolescents: Evidence for protracted development of speech motor control processes. *Journal of Speech, Language, and Hearing Research, 45*(6), 1119-1133.

Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49*(1), 25-47.

Wohlert, A. B., & Smith, A. (2002). Developmental change in variability of lip muscle activity during speech. *Journal of Speech, Language, and Hearing Research, 45*(6), 1077-1087.

Wong, P., Schwartz, R. G., & Jenkins, J. J. (2005). Perception and production of lexical tones by 3-year-old, Mandarin-speaking children. *Journal of Speech, Language, and Hearing Research, 48*(5), 1065-1079.

Xu, C. X., & Xu, Y. (2004). Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association, 33*(02), 165-181.

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

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics, 27*(1), 55-105.

Xu, Y. (2001). Sources of tonal variations in connected speech. *Journal of Chinese Linguistics, monograph series #17*, 1-31.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication, 46*, 220-251.

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America, 111*, 1399-1413.

Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication, 33*(4), 319-337.

Xu, Y. (2008). *TimeNormalizeF0.praat* (Version 2.6.4) [Computer script] Retrieved June 26, 2008, Available from http://www.phon.ucl.ac.uk/home/yi/tools.html

Yip, M. (2002). *Tone* (first ed.). United Kingdom: Cambridge University Press.

Zemlin, W. R. (1988). *Speech and hearing sciences: Anatomy and physiology* (3rd ed.). Jersey: Prentice-Hall, Inc.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool language scale, fourth edition (PLS-4) English edition* (fourth ed.). San Antonio, Texas: Harcourt Assessment, Inc.