Effects of variance and input distribution on the training of L2 learners' tone categorization

By

Jiang Liu

Submitted to the graduate degree program in Linguistics and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chairperson: Jie Zhang

_____

Allard Jongman

_____

Joan Sereno

_____

Annie Tremblay

_____

David Johnson

Date Defended: 08/30/2013

The Dissertation Committee for Jiang Liu

certifies that this is the approved version of the following dissertation:

Effects of variance and input distribution on the training of L2 learners' tone categorization

_____

Chairperson: Jie Zhang

Date approved: 08/30/2013

Abstract

Recent psycholinguistic findings showed that (a) a multi-modal phonetic training paradigm that encodes visual, interactive information is more effective in training L2 learners' perception of novel categories, (b) decreasing the acoustic variance of a phonetic dimension allows the learners to more effectively shift the perceptual weight towards this dimension, and (c) using an implicit word learning task in which the words are contrasted with different lexical tones improves naïve listeners' categorization of Mandarin Chinese tones. This dissertation investigates the effectiveness of video game training, variance manipulation and high variability training in the context of implicit word learning, in which American English speakers without any tone language experience learn four Mandarin Chinese tones by playing a video game. A video game was created in which each of four different animals is associated with a Chinese tone. The task for the participants is to select each animal's favorite food to feed it. At the beginning of the game, each animal is clearly visible. As the game progresses, the images of the animals become more and more vague and eventually visually indistinguishable. However, the four Chinese tones associated with the animals are played all through the game. Thus, the participants need to depend on the auditory information in order to clear the difficult levels. In terms of the training stimuli, the tone tokens were manipulated to have a greater variance on the pitch height dimension, but a smaller variance on the pitch direction dimension, in order to shift the English listeners' perception to pitch direction, a dimension that native Chinese listeners crucially rely on. A variety of pretests and posttests were used to investigate both the English speakers' perception of the tones and their weighting of the acoustic dimensions. These training stimuli were compared to other types of training stimuli used in the literature, such as the high variability natural stimuli and tones embedded in non-minimal pairs. A group of native English speakers was used as the control group without any tone input. A native control group was also included. The video game training for each speaker consisted of four 30-minute sessions on four

different days, and 60 participants (including both the non-native control and native control group) participated in the experiments.

The crucial findings in the study include (1) all naïve listeners in the training condition successfully associated lexical tones with different animals without any explicit feedback after only 2 hours of training; (2) both the resynthesized stimuli with smaller variance on pitch direction and the multi-talker stimuli allowed native English speakers to shift their cue-weighting toward pitch direction and the multi-talker stimuli were more robust in terms of shifting the cue-weighting despite their more heterogeneous distribution in the acoustic space; (3) the multi-talker training allowed for better generalization as the trainees in multi-talker training identified the tones produced by new talkers better than trainees in other conditions; (4) there was a main effect of tone on tone identification and the falling tone was the most challenging one; (5) there is a correlation between cue-weighting and the tone discrimination performance before and after the training; (6) due to individual variability, individuals differed in terms of the amount of tone input they received during the video game training and the number of tone tokens was a significant predictor for the sensitivity to tones calculated as d'. Overall, the study showed an effect of talker variability and variances of multidimensional acoustic space on English speakers' cue-weighting for tone perception and their tone categorization.

Acknowledgments

The very first person I would like to thank is Prof. Jie Zhang, the chair of my dissertation, the mentor for my academic study and the friend of my life. No matter how "harsh" he was in training me to be a linguist, I feel lucky to have an academic advisor like him who gives me a hard time but really makes me learn things. I learned so much from him. I not only learned how to do linguistic research from Prof. Zhang but also inherited his passion and enthusiasm for studying linguistics. He really is a role model for me to be a skilled researcher, a good time manager, a good communicator and a person who has faith in his research and career. Looking back on my years in graduate school, Prof. Zhang not only transformed me academically but more importantly shaped my character. I used to be scared of being asked questions because I was afraid of not being able to answer them. But now I really like taking questions. No matter whether I can address them or not, I can always delve deeper into my research through these questions. I also learned how to defend my viewpoint through countless brainstorming sessions with Prof. Zhang. No matter whether I eventually hold my viewpoint or give it up, it makes me think really hard on serious issues in phonology and phonetics. He is a phonologist, but he is open minded about the research topics I chose. It eventually made me find the research area I am really passionate about. Though he constantly challenges my research ideas, he never discourages me, and the references he provided me helped me greatly in developing my research ideas. Prof. Zhang has been supportive of me from the very first day of my graduate career to the end of my PhD study. He never gave up on me even when I myself started to lose faith in the research direction I was pursuing. Without his full support, I will not be able to get the research funding for my dissertation research. Without all these years' training and working with Prof. Zhang, I will not be able to develop research ideas that make serious inquiries about phonetics and phonology, let alone to do a dissertation.

Two other people I would like to express my special thanks to are Prof. Allard Jongman and Prof. Joan Sereno. They have been keeping a close eye on my academic progress during my graduate study. Whenever I need help, they are always available even

when they were on sabbatical. They are always generous in terms of spending time talking to me, even teaching me how to produce publishable graphs. They are really supportive. Prof. Jongman provided me with the technical support whenever I was in need. Prof. Sereno helped me coordinate with other professors in terms of allowing me to get things done in time. I really appreciate the trust that Prof. Sereno endowed me to run speech perception experiments with a large scale for her. Such experience helped me be able to run experiments with a large number of participants for my dissertation research within a relatively short period of time. In a way, my dissertation topic was developed partially due to their wide connection to psychologists who are interested in language. It really broadened my horizon in the research area of phonetics and psycholinguistics.

My other dissertation committee members Prof. Annie Tremblay and Prof. David Johnson also provided tremendous help to my dissertation. Taking the seminar course taught by Prof. Tremblay really made me delve into the research area of speech production and perception. Prof. Tremblay is also very supportive by helping me find participants for my research and answering my questions regarding experimental design. Prof. Johnson taught me INDSCAL, a statistical tool that is widely used for studying tone perception and indispensable for my dissertation research. I am deeply indebted to all of my committee members. I really appreciate their support and help, which I may not be able to pay back for life.

I also want to express my enormous gratitude to all my fellow students in the Linguistics Department at KU. We are Jayhawks who really support each other and care for each other. It was really fun studying, working and playing with all my fellow students. Working with Hyunjung Lee let me know what is ‘diligence’. Working with Steve Politzer-Ahles let me know what is ‘erudite’. The snacks and drinks Goun Lee provided in the lab down the basement made me feel it is a cozy place to stay rather than a cell without sunshine.

I also want to thank Prof. Yan Li and Yue Pan in the East Asian Languages and Cultures Department at KU who helped me enormously as well. They taught me how to teach Chinese as a second language and it opened another path for my career.

I am deeply indebted to my parents, my wife and parents in law. Though knowing my interests in language may not be able to lead me to have a life of abundance, my parents were never against the major I chose and always provided me with whatever they could to support my study. I just want to say “Thank you, mom and dad.” (感谢父母). I am also indebted to my wife. In spite of her popularity among male students in medical school, she eventually made the ‘wrong decision’ to marry me. She gave up her practice in China and came to the U.S. for graduate study in basic research because of me. I am overwhelmed by the love she gives me. With her love, I never feel alone. Finally, I want to thank my parents-in-law who did not stop my wife from being with me and showed lots of care about my health. Without their support, I will not be able to have a happy family.

Last but not least, I want to thank Dr. Yao Zhao and Mr. Shengcai Yang who helped me create the video game for my dissertation research. The life-long friendship between Yao and me shall never vanish.

Table of Content

List of Figures

List of tables

CHAPTER ONE: INTRODUCTION

Humans perceive speech categorically. Given a speech stimulus continuum, listeners identify a group of continuum steps either as one sound category or another. Within category discrimination is usually much poorer than cross-category discrimination (see Strange 1995 for an overview of categorical perception of segments; see Hallé et al. 2004 and Xi et al. 2010 for categorical perception of Mandarin Chinese tones). In terms of L2 sound categorization, non-native speakers do not always perceive L2 sound categories in the same way as the native speakers do. For example, Japanese speakers depend primarily on F2 rather than F3 for distinguishing English /r/ and /l/ whereas English speakers depend primarily on F3 for the /r/ and /l/ distinction (Yamada 1995, Iverson et al. 2003). For tone perception, American English speakers depend more on pitch height (average pitch) whereas native speakers of Mandarin Chinese depend more on pitch slope (Gandour 1983, Huang 2001). Based on these cross-linguistic perception studies, researchers have developed phonetic training paradigms that aim at training L2 learners to have more nativelike perception (e.g., Bradlow et al. 1997, Iverson et al. 2005, Wang et al. 1999, Wong and Perrachione 2007, Goudbeek et al. 2008).This body of research has shown evidence of plasticity in the adult system to support non-native sound category learning even though the system is not as flexible as in earlier development (e.g., Bradlow et al. 1997; Goudbeek et al. 2008).

Several phonetic training paradigms were developed in the past mainly for the purpose of helping native Japanese speakers' categorization of L2 English /r/ and /l/. Previous studies have found that Japanese English learners had enormous difficulty distinguishing L2 English /r/ and /l/ because they used F2 as the primary acoustic cue for categorizing L2 English /r/ and /l/ whereas native English speakers used F3 as the primary acoustic cue (Lively et al. 1993, Iverson et al. 2003). Acoustically, F3 is a more robust acoustic cue for distinguishing English /r/ and /l/ relative to F2 because /r/ and /l/ have a larger overlap on F2 (Iverson et al. 2003, Lotto et al. 2004).

Strange and Dittmann (1984) trained Japanese listeners in a discrimination paradigm with a synthetic "rock"-"lock" stimulus continuum. They assumed that training listeners with tokens that contrasted /r/ and /1/ in initial singleton position would allow subjects to form a prototype that could be applied to other phonetic environments. However, the results showed that only a limited number of subjects improved their /r/-/l/ discrimination for the natural stimuli and only one subject improved the discrimination at other non-initial positions. The result suggested that discrimination training with a small set of tokens from one phonetic environment may be ineffective in modifying listeners' phonetic perception for L2 sound categories. Logan et al. (1991) developed a High Variability Phonetic Training (HVPT), which involves having subjects give identification judgments with feedback for natural recordings of words produced by multiple talkers, with target phonemes in multiple syllable positions. The training result of HVPT turned out to be very robust in terms of generalizing /r/-/l/ identification to new talker and new contexts. Iverson et al (2005) studied the effectiveness of both HVPT and resynthesized

stimuli for Japanese speakers' identification of L2 English /r/ and /l/. The result showed that both the manipulation of F2 and F3 in /r/ vs. /l/ and the multi-talker training significantly improved Japanese speakers' identification of /r/ and /l/; however, the resynthesized stimuli did not achieve better training result than the multi-talker training. The robustness of multi-talker training had been used as strong evidence for certain psychological models on sound categorization such as the exemplar model (e.g., Nosofsky 1986; Kruschke 1992). We will provide more details about different models on sound categorization in Chapter Two.

More recently, several psycholinguistic studies found that by manipulating the variance on different acoustic dimensions, it is possible to shift listeners' cue-weighting from one acoustic dimension to another within a relatively short training period (e.g., Holt and Lotto 2006). The method of variance manipulation had been applied to the training of Japanese listeners' perception of L2 English /r/ and /l/ (Lim and Holt 2011). The result showed that making the variance on F2 larger than the one on F3 helped Japanese speakers shift cue-weights towards F3, which is the primary acoustic cue native English speakers use for /r/ and /l/ distinction. With more cue-weighting shifted towards F3, the participants' identification of /r/ and /l/ also significantly improved. The phonetic training paradigms developed during the past decades not only have been proved to be helpful for improving L2 learners' perception of L2 sound categories, but also shed light on the nature of human sound categorization.

Apart from the difficulty of segments like /r/ and /l/ raised in L2 learning, suprasegmental features such as Mandarin Chinese tones can also be challenging in L2

learning. By definition, languages that exploit variations in pitch to differentiate word meanings are called tone languages (Yip 2002). Mandarin Chinese, a tone language, uses four tones, as exemplified by: ma1 'mother' [T1: high level], ma2 'hemp' [T2: high rising], ma3 'horse' [T3: dipping], ma4 'scold' [T4: high falling]. Non-tone language speakers who do not use tones to contrast word meanings in their native languages often have more difficulty learning L2 Chinese tones than novel L2 segmental units at the beginning stage. Thus, it is worthwhile to develop a phonetic training paradigm that can help non-tone language speakers form lexical tone categories efficiently.

Based on the advances in the study of human's speech perception such as cue-weighting of different acoustic cues in sound categorization and methods for L2 speech learning, the current study used a state-of-the-art phonetic training paradigm, a multi-modal phonetic training paradigm, to further study humans' speech perception and examine the robustness of the different types of training stimuli in improving L2 sound categorization. We aimed to investigate what type of information in the training input is the most useful and efficient in terms of forming L2 tone categories, using Mandarin Chinese lexical tones as the target sound categories. In order to avoid lexical effects, neighborhood density, word frequency and other language-specific factors that may contribute to the acquisition of L2 sound categories, we used the syllable /y/ (rounded version of the vowel /i/) that carries four different lexical tones, which are completely novel to native English speakers, as the stimuli to study L2 tone categorization of naive listeners (native English speakers without any tone language experience in this case). Applying the phonetic training paradigms, which had been used for the perceptual

training of L2 segmental units (e.g., the consonants /r/ and /l/), to the perceptual training of suprasegmental units such as lexical tones can inform us whether the training paradigms are also useful for L2 tone categorization. In particular, comparing training stimuli in which the variances on relevant acoustic dimensions are manipulated with multi-talker training stimuli can provide us with a better understanding of what the causes for cue-weighting shifts are. An earlier tone training study has shown that it is possible to shift English speakers' cue-weighting for tone perception (Chandrasekaran et al. 2010). However, no study has examined whether the more nativelike cue-weighting has a significant impact on the discrimination of specific tone pairs (e.g., high level tone vs. low dipping tone; rising tone vs. falling tone). Thus, in the current study, we examined both the English speakers' cue-weighting and the discrimination for different tone pairs before and after the training. In practice, previous studies on cross-linguistic tone perception mostly used Reaction Time (RT) to study the participants' cue-weighting of tone perception (Gandour 1983, Francis et al 2008), the current study examined both RT and accuracy of the discrimination of different tone pairs in order to see if there is a correlation between cue-weighting and discrimination accuracy for different tone pairs. The basic research on L2 tone categorization also has significant implications for L2 Chinese teaching, particularly, the training of listening comprehension, since L2 Chinese learners at the beginning stage have difficulty discriminating and identifying the four Chinese lexical tones.

The dissertation is organized as follows. Chapter Two provides the background on sound categorization, the effect of talker variability and variances in the acoustic space on

cue-weighting for sound categorization, phonetic training for L2 sound categorization, the effect of input distribution on sound categorization, and phonetic training for L2 Chinese tone learning. The research questions and hypotheses of the current study are also set up in Chapter Two. Chapter Three provides the details of the methods and experimental design of the current study. Chapter Four reports the results of all the experiments conducted for the learning of L2 Chinese tones. Chapter Five discusses the results and the related theoretical implications. Chapter Six is the conclusion.

CHAPTER TWO: BACKGROUND AND RESEARCH QUESTIONS

2.1 Theories of sound categorization and the effect of talker variability on sound categorization

Early psychological research on vision led to fruitful theoretical models about visual categorization. Three distinct theories of visual categorization have featured most prominently in the recent literature. The decision-boundary theory (Ashby & Perrin, 1988) assumes that categorization is based on the comparison of the perceptual effect of a stimulus with category boundaries stored in memory. The prototype theory (Rosch, 1973) assumes that stimuli are categorized on the basis of their similarity to category prototypes stored in memory. A category prototype is generally defined as the average, or the most typical, member of a category. Finally, the exemplar theory (Nosofsky, 1986), conversely, denies the explicit use of category prototypes. In its extreme formulation, exemplar theory assumes that categorization is based on a comparison of the stimulus with all previously categorized exemplars of all categories.

These models on the categorization of visual stimuli had shaped the research trajectory for sound categorization in the last century. Psychological evidence for the decision-boundary model came from the finding that the identification of sound categories depends on the noise involved in the sound categories. With a speech continuum, depending on the physical distance between the steps, the smaller the distance is, the larger the noise becomes, and the larger the distance is, the smaller the noise

becomes. The categorization slope can be influenced by the noise involved in the test stimuli. It is because the probability of assigning the labels to the stimuli becomes similar as they are obscured by the noise (Ashby & Perrin, 1988).

Psychological evidence for phonetic prototypes comes from several sources (see Kuhl 1991a). First, some members of a phonetic category are responded to faster, more accurately, and with higher confidence ratings or goodness judgments than others (see Kuhl 1991a, b). Second, some within-category discriminations are better than others. For example, Kuhl found that tokens of /i/ surrounding a good exemplar were more poorly discriminated than tokens of /i/ that surrounded a poor category exemplar (Kuhl 1991 a,b; Kuhl et al., 1992). Finally, Miller (1977) and Repp (1976) demonstrated that good category members are more effective competitors in dichotic listening conditions than poor category members. In addition, from the perspective of formal linguistic analyses, sound categories are abstract, which are context invariant. The sound system consisted of phonemes whose phonetic variants are grouped together based on complementary distribution, free variation or phonetic similarity. From the formal linguistic point of view, the mental representation of sounds is abstract and therefore fits the prototype model in practice. Computational simulations of sound categorization using the prototype model have also been conducted (see Liljencrants & Lindblom 1972; de Boer 2000).

Since 1990s, Exemplar model received a substantial amount of support from many psycholinguistic studies such as phonetic training studies on L2 sound categorization (e.g., Lively et al 1993; Logan et al 1991) and child language acquisition studies (e.g., Rost and McMurray 2009). For example, Lively et al. (1993) trained the perception of L2 English /r/ and /l/ on two Japanese speaker groups, one of which was trained with a single speaker's stimuli and the other was trained with multiple speakers'

stimuli. There were two crucial findings. The first finding was that the multi-talker training allowed Japanese speakers' identification of /r/ and /l/ to generalize to new talkers and new phonetic environments whereas the single-talker training group failed to make such generalizations. Second, effects of talker variability were observed throughout multi-talker training. Accuracy and response latency varied widely as a function of talker in the training. These findings suggested two conclusions. First, the improvements obtained during the training reflect stimulus-specific learning, rather than robust abstract category acquisition. True category acquisition would be demonstrated by generalization to new talkers and new tokens over many different environments. Second, the presence of talker variability in the stimulus set during training appears to be an important condition for demonstrating robust generalization in this type of training paradigm.

The effect of talker variability for improving sound categorization was not only observed among adults but also observed in child language acquisition. Rost and McMurray (2009) conducted a study that trained infants to learn minimal word pair /puk/ vs. /buk/ by using either single talker stimuli or multi-talker stimuli. The 14-month old infants were divided into two groups, one of which was exposed to exemplars of the minimal pair produced by a single speaker and one of which was exposed to exemplars of the minimal pair produced by multiple speakers. During the training, pictures were accompanied by the two sound labels. After reaching habituation (a criterion of familiarization of the picture and sound), in the test phase, the infants were presented with the picture that was accompanied by either the correct sound label (the match condition) or the wrong sound label (mismatch condition). Only the multi-talker training group showed a significant longer looking time in the mismatch condition relative to the match condition (a longer looking time suggests that the infant is able to discriminate the minimal pairs) whereas the single-speaker training condition did not show any looking

time difference between the two conditions (the same looking time suggests not being able to discriminate the minimal pairs). The conclusion based on this result was that lexical neighbor learning could be improved by incorporating greater acoustic variability in the words being learned, as this may buttress the still-developing phonetic categories and help the infants identify the relevant contrastive dimensions. In this case, the effect of talker variability was proved to be robust for word learning, which is the ultimate goal of sound categorization.

An important assumption made in the exemplar model is that subjects store in memory multidimensional representations of objects presented during training. A selective attention mechanism weighs the importance of various stimulus dimensions. The critical dimensions for category membership are given strong weights, while dimensions that are less important receive smaller weights. Changes in selective attention "stretch" and "shrink" the perceptual space for these dimensions and in turn alter the internal category structure: Objects become less similar to each other as dimensions are stretched and more similar to each other as dimensions are shrunk (Nosofsky 1986; Kruschke 1992). From the results presented in Lively et al (1993), the researchers argued that listeners encoded talker-specific information in memory. One consequence of the high variability multi-talker training was that representations for the new phonetic categories were stretched on certain relevant acoustic dimensions given different voices. As a result, subjects had a relatively unconstrained set of exemplars from which to generalize. Listeners in single talker training, in contrast, were trained with a highly constrained stimulus set and as a consequence showed poor generalization to a new talker. It appeared that training with a single talker was a relatively ineffective method for stretching listeners' perceptual space for non-native contrasts. Rather, subjects engaged in stimulus-specific category learning. Similarly, for Rost and McMurray (2009), to account

for the robustness of the multi-talker training for learning minimal pairs, the authors argued that there were at least two kinds of relevant variability that may be important for learning minimal pairs. One is the variability along specifically phonetic dimensions (e.g., VOT). The other is the variability in non-phonetic information (e.g., voice quality, gender), which may help learners extract the relatively invariant phonetic dimensions. The first type of variability plays a crucial role in allowing the learners to define the phonetic or lexical categories that contrast the words.

Though toward the end of the first year of life, there are indications that the sensitivity to some speech contrasts that do not appear in the native language begins to decline (Werker and Tees 1984a), as Lively et al. (1993) and several other studies (e.g., Logan et al 1989; Pisoni et al. 1982) showed, the decline is not a permanent one because listeners can be retrained to perceive such distinctions. Even without training, previous study showed that both adults and 12- to 14-month-old infants from English-speaking homes were able to discriminate Zulu click contrasts. There was no indication of any decreased sensitivity to these contrasts despite the fact that they do not occur in English (Best et al. 1988). It is possible that the acoustic dimension that is relevant for the click contrasts is a completely new dimension to native English speakers. The new acoustic dimension does not interfere with the acoustic dimensions already used for the English sound contrasts. Once selective attention was drawn to the dimension, it imposes no difficulty discriminating the new sound categories. Consequently, one interpretation of these results is that the declines observed are attributable to shifts in attention away from the dimensions that distinguish the foreign language contrasts and the reorganization of the perceptual space of the sounds is still possible.

We want to highlight another study that used multi-talker training for Japanese speakers' perception of L2 English /r/ and /l/ here because it is a study that comprehensively examined the efficiency of using resynthesized stimuli and multi-talker stimuli for L2 English /r/ and /l/ categorization. Iverson et al. (2005) manipulated F2 and F3 in L2 English /r/ and /l/ (e.g., amplifying F3 difference between the two sound categories, reducing F2 difference) in order to examine the effect of the resynthesized stimuli for Japanese speakers' categorization of L2 English /r/ and /l/. They also used HVPT (High Variability Phonetic Training) as the baseline. The four training paradigms in Iverson et al. (2005) are summarized in Table 1:

Table 1. Four training paradigms used in Iverson et al. (2005)

Training paradigm	**Training data**
HVPT (High Variability Phonetic Training) — natural stimuli	Identification of English /r/ and /l/ at different syllable positions with feedback by using multi-talkers' tokens.
All enhanced training — cue-manipulated stimuli	Extreme F3 values during closure for /r/ and /l/: 100Hz higher than median F2 for /r/ 100Hz lower than median F4 for /l/ F3 enhancement reduced linearly during transition from closure to vowel. Back to original F3 at the end. Same stimuli were used from Day 1 to Day 10.
Perceptual fading — cue-manipulated stimuli	F3 difference between /r/ and /l/ decreased from Day 1 to Day 10. First F3 was set to extreme values and then gradually reduced.
Secondary cue variability — cue-manipulated stimuli	F2 difference between /r/ and /l/ gradually increased throughout training period. Day 1: no F2 difference (F2 set to median F2 for both /r/ and /l/. Day 10: maximum and minimum F2 values were used but randomly combined with short and long closures and transitions.

The crucial finding of Iverson et al. (2005) was that all training conditions significantly improved Japanese speakers' L2 English /r/-/l/ distinction. However, there was no significant difference across training conditions in terms of post-test /r/-/l/ identification accuracy rates. The authors also reported that there was a weaker generalization in terms of the identification of /r/ and /l/ produced by new talkers or at new syllable positions. The authors concluded that cue-manipulated training paradigms did not improve the /r/-/l/ categorization above HVPT. In terms of cue weighting, the authors calculated the d' and bias for different acoustic parameters. Only bias result was reported. The bias statistics provided a way of measuring cue weighting. For example, if the listeners were biased to identify stimuli with long transitions as /r/ and short transitions as /l/, this would demonstrate that the transition duration affected whether they identified the stimulus as /r/ or /l/ and thus indicate that transition duration had a high weighting in the categorization decision. If listeners had zero bias for a cue, this would indicate that the cue did not affect /r/-/l/ identification, and thus had low weighting. Surprisingly, after the training, the HVPT, the perceptual fading and the secondary cue variability training did not lead learners to have a higher cue weighting on F3. Only the All Enhanced training condition allowed the listeners to have a small cue-weighting increase on F3. The results suggested that both the resynthesized stimuli and the multi-talker training stimuli were able to improve Japanese speakers' categorization of L2 English /r/ and /l/, but only one certain type of resynthesized stimuli was capable of shifting Japanese speakers' cue-weighting to be more nativelike. However, several studies that used multi-talker phonetic training indeed showed the effect of shifting

learners' cue-weighting to be more nativelike (e.g., Wong and Perrachione 2007; Chandrasekaran et al. 2010). As Iverson et al (2005) showed, setting a large difference between /r/ and /l/ on the relevant acoustic dimension (i.e., F3) can make the cue-weighting increase on this dimension. More recently, researchers found that manipulating the variance on different acoustic dimensions in the resynthesized stimuli was quite effective in terms of shifting cue-weighting to be more nativelike for Japanese speakers' categorization of L2 English /r/ and /l/ (Lim and Holt 2011). It is unclear why Iverson et al. (2005) did not find a more nativelike cue-weighting shift for multi-talker training. It is still reasonable to believe the existence of multi-talkers' training effect on cue-weighting shifts as Pisoni and colleagues (e.g., Pisoni et al 1994; Lively et al. 1993) have argued that exposing listeners to a wide range of talkers' stimuli is better than training with a small range of talkers' stimuli because the distributions of natural stimuli teach learners which cues are the most reliable; listeners are thought to store individual exemplars that they hear in training, and the multidimensional categorization space for these stimuli gets stretched along dimensions where sound categories differ and get shrunk along dimensions that do not distinguish the sound categories (see Nosofsky 1986).

The current study hypothesizes that, according to the exemplar theory, multi-talker training is not only able to improve the performance on tone categorization but also has the effect of shifting cue-weighting for tone perception as well. Another reason for our preference for the exemplar model over the prototype model is that the distinct sound categories may not be normally distributed in the multidimensional acoustic space, in which case it is possible that even good exemplars may be far away from the centroid of

the exemplars/prototype. Thus, merely comparing the input signal to the centroid may cause errors more easily. In Section 2.2, we review studies that manipulated the variance on different acoustic dimensions for training of L2 sound categorization and the effect of variance on shifting cue-weighting.

2.2 Effect of variance on sound categorization and multi-modal phonetic training

Several recent psycholinguistic findings shed light on L2 phonetic training. Smits, Sereno and Jongman (2006) systematically tested the decision boundary model, prototype model and distributional/exemplar model for sound categorization by making native Dutch listeners learn two non-speech categories where the mean and variance were manipulated either on the formant dimension or the duration dimension orthogonally. They trained the listeners with two sets of non-speech stimuli, which belonged to two different categories. Feedback was provided during the learning session for the participants to learn two distinct categories. They manipulated the distance between two categories' means and the standard deviations of the two categories in terms of just noticeable difference (jnd) units. One innovative part of this study is that the researchers used a synthesized continuum either on the formant or duration dimension as the test stimuli in a sound identification task after the training. Using the synthesized continuum as the test stimuli can help obtain the categorization slope for each individual. With mathematical formulation, the three sound categorization models make different

predictions about the relationship between the distribution of the training stimuli and the ultimate categorization slope. The decision boundary model predicts that as long as an optimal boundary between two categories can be obtained during the training the means and standard deviations of the training stimuli will not affect the categorization slope. The only thing that matters is the noise in the test stimuli. The larger the distance between the continuum steps next to each other, the larger the noise is. It makes the categorization slope shallower. In other words, different sets of training stimuli with different means and standard deviations will not change the categorization slopes. The prototype model predicts that only the distance between the means of the two sound categories affect the categorization slopes. The closer two categories' means are the shallower the categorization slope is. Finally, the exemplar model/distributional model predicts that both the means and standard deviations affect the categorization slope. Given the distance between two sound categories' means, the categorization slope is proportional to the standard deviation of the two categories (in this case, the standard deviations of two categories were the same).

Interestingly, the results showed that the decision boundary model was supported and the exemplar/distributional model was partially supported by the sound categorization result on the formant dimension. However, the sound categorization result on the duration dimension supported both the decision boundary model and the exemplar/distributional model. In their study, the only theory that remains unsupported is the prototype theory. As the researchers argued, particular versions of the decision boundary model also need the distribution information (e.g., means and standard

deviation) of the training stimuli to draw the boundary between distinct sound categories. Thus, overall, the distribution information is indispensible for establishing distinct sound categories.

The results also suggested that different categorization mechanisms seem to operate for the two acoustic dimensions (i.e., formant and duration). The researchers argued that some psychophysics theory may provide an explanation. Stevens and Galanter (1957) introduced the concepts of prothetic and metathetic scales. A prothetic scale is a psychological scale to which, at a physiological level, an "additive" mechanism applies—that is, increasing a value on a prothetic scale is equivalent to adding more of the same. Examples of prothetic scales are brightness, loudness, and duration. A longer sound simply has "more duration" than a shorter sound and is presumably encoded at a physiological level by a stronger or longer firing of basically the same neurons. In contrast, a "substitutive" mechanism applies for metathetic scales, such as (visual) position, pitch, and, presumably, timbrelike magnitudes, such as formant frequency. A pure tone with a higher pitch does not simply have "more frequency" than one of a lower pitch. Instead, it essentially stimulates different fibers in the auditory nerve. Empirically, the difference between the two scales is evidenced by the fact that for metathetic scales, the jnd measured in subjective units is constant across the scale (e.g., the jnd for pitch expressed in mels is the same for low and high tones), whereas the same does not hold for prothetic scales (the jnd for loudness expressed in sones is smaller at the low end of the scale than at the high end). If different acoustic cues are psychophysically encoded differently then the conclusions in Smits et al. (2006) may not be easily generalized to

sound categorization where acoustic cues other than formant or duration are included in the input. Nevertheless, Smits et al (2006) is one of the few pioneer studies that looked into the relationship between the distribution of the input sound stimuli and the ultimate sound categorization.

Goudbeek et al. (2005) is another study that examined the effect of the distribution on multiple acoustic dimensions on sound categorization. They compared adult American English speakers' learning of non-speech categories and phonetic categories by manipulating the distributions of the training stimuli on two acoustic dimensions simultaneously. In the non-speech condition, American English speakers categorized stimuli that simultaneously varied in duration and resonant frequency. Two conditions were created. The first condition was that the two non-speech categories can be distinguished just using the duration dimension. The second condition was that the two non-speech categories can be distinguished only if both duration and resonant frequency are taken into account. The test stimuli were synthesized continuum with equal steps either on the duration dimension or on the resonant frequency dimension. With the test stimuli, the cue-weighting on each dimension can be calculated by logistic regression. The results showed that, for the first training condition, the listeners can easily ignore the spectral information in both training and test when only the duration can be used to distinguish the two non-speech categories. For the second condition where both duration and resonance frequency need to be used for distinguishing two sound categories, six out of twelve subjects used both dimensions in the training but only one out of twelve subjects used both dimensions in the test phase. Most subjects still heavily relied on the

duration dimension to categorize the non-speech sounds. In the second experiment of learning phonetic categories (i.e., three synthesized Dutch front vowels), a similar pattern was found. The difference among the three Dutch vowels can be described by duration and the first formant as well.

Regardless of the relevance of two acoustic dimensions for distinguishing different sound categories, it seemed that the learners cannot learn to use both dimensions. There are two possible reasons for this. One is that the duration cue was always used as the only relevant acoustic cue in the experiment before the experiment in which both duration and first formant were used as acoustic cues. Then it may bias the learners toward duration. The second possible reason is related to the first one. If the listeners were biased toward the duration cue, the equal variances on the duration dimension and the formant dimension may not be able to shift listeners' cue-weighting from the duration to the formant. Holt and Lotto (2006) showed how the manipulation of variance on different acoustic dimensions may adjust cue-weighting for non-speech categorization. In their study, they synthesized two non-speech categories characterized by Center Frequency (CF) and Modulated Frequency (MF). By decreasing the variance on MF dimension and increasing the variance on CF dimension, listeners shifted their cue-weighting from CF towards MF.

Lim and Holt (2011) utilized the idea of variance manipulation to train Japanese listeners' identification of L2 English /r/ and /l/ by decreasing the variance on F3 and increasing the variance on F2. The result showed that cue-weighting on F3 indeed increased in the post-tests and the identification accuracy rate improved significantly. To

control the potential effect that the video game alone can improve Japanese listeners' perception of /r/ and /l/, the researchers included a control condition in which only synthesized non-speech stimuli were used in the video game training. Since the participants in the control condition did not show any improvement for identification of /r/ and /l/, it excluded the possibility that it is the game alone that helped improve the categorization of /r/ and /l/. In their training, they integrated the variance-manipulated exemplars of /r/ and /l/ with a 3D alien shooting game. Either a /ra/ or /la/ exemplar was played repeatedly depending on the specific type of alien that appeared on the screen. For the alien accompanying the /ra/ sound, the subjects needed to shoot at it whereas for the alien accompanying the /la/ sound, the subjects needed to capture it. At first, the shape and color of the aliens were clearly visible. But as the game progressed, the shape and color became more and more vague. Thus, the subjects needed to depend more on the sounds that accompanied the alien in order to decide whether to shoot or capture it. The researchers argued that training Japanese speakers' perception of /r/ and /l/ through such an implicit sound category learning paradigm in a multi-modal interface made the learning more efficient.

Comparing the Japanese speakers' post-test identification accuracy rate in Lim and Holt (2011) with previous phonetic training studies on Japanese speakers' English /r/-/l/ categorization (e.g., Bradlow et al. 1997, Iverson et al. 2005), we find that the identification accuracy rates are similar among these studies (about 80%), however, the training period in Lim and Holt (2011) is considerably shorter. In Lim and Holt (2011), the total training period only lasted 2.5 hours compared to 5 hours in Iverson et al. (2005)

and 4 weeks (10 hours in total) in Bradlow et al. (1997). As Lim and Holt argued, the comparable learning result, but with a much shorter training time in the video game training suggests the possibility that functional use of sounds may facilitate complex, multidimensional category learning even without an overt categorization task. However, without an experiment that uses an overt categorization task and the same variance manipulated training stimuli, it is unclear whether it is the video game or the variance manipulated training stimuli that are responsible for the higher training efficiency.

Although we may not be able to make a strong claim about the robustness of video game training based on Lim and Holt's (2011) results, some research has shown that videogames produce robust perceptual learning (e.g., Green & Bavelier, 2007) and may be highly effective at activating the striatal reward system of the brain (e.g., Koepp et al., 1998), providing the intrinsic learning signals. Based on human neural imaging research, Tricomi et al. (2006) propose that tasks that include goal-directed action for which there is a positive or negative outcome contingent on one's behavior, tasks in which actions are performed in the context of expectations about outcomes, and tasks in which individuals have incentive to perform well (Delgado, Stenger, & Fiez, 2004) are most likely to robustly activate striatal reward system processing (within the caudate nuclei, in particular). Recruitment of the striatal reward system may facilitate learning through feedback to perceptual representations and, additionally, may affect learning through its influence on other mechanisms, such as Hebbian learning (Vallabha & McClelland, 2007), by serving as an informative signal to guide learners to better differentiate available information (Callan et al., 2003). Other research has also

demonstrated that learning can be driven by extrinsic rewards like performance feedback or monetary gains (Seitz et al. 2009) or by intrinsically generated performance evaluation in lieu of feedback (Seitz & Watanabe, 2009). The videogame paradigm in Lim and Holt (2011) is unique in that it lies between these two endpoints. Participants did not engage in explicit categorization and did not receive performance feedback about categorization. Yet learning is not entirely passive or unsupervised; feedback arrived in the form of success or failure in achieving one's goals in the game and there are multiple, correlated multimodal events and objects that co-vary with speech category membership. These characteristics may engage intrinsically generated learning signals to a greater extent than passive training paradigms and, perhaps, even to a greater degree than extrinsic performance feedback. In support of intrinsic learning, Wade and Holt (2005) found that the non-speech auditory category learning within the game exceeded unsupervised learning of the same sounds. In support of passive learning, direct attention to target stimuli can sometimes actually hamper perceptual learning (Tsushima, et al. 2008). Lim and Holt (2011) argued that the videogame paradigm is intrinsic learning in nature. The relatively high motivation and engagement elicited by video-games (especially compared to standard, overt categorization tasks) may evoke greater intrinsic reward-based processing supportive of learning. The intrinsic reward of success in the game, of accurately predicting and acting upon upcoming events, may be a powerful signal to drive learning.

In sum, Lim and Holt (2011) showed that (1) video game phonetic training increases learning efficiency of L2 sound categorization, and (2) variance manipulation

phonetic training leads to cue-weighting shift towards the acoustic dimension on which the sound categories have a smaller variance.

2.3 Input distribution effect on sound categorization

Early studies have shown that infants learn about the phonetic categories of their language between six and twelve months, meanwhile gradually lose their sensitivity to non-native contrasts (e.g., Werker & Tees, 1984). The learning mechanism that accounts for the phonetic category formation is known as distributional learning (Maye, Werker, & Gerken, 2002). This account proposes that learners obtain information about which sounds are contrastive in their native language by attending to the distributions of speech sounds in the acoustic space. In their study, learners tended to infer two categories when given a bimodal distribution of sounds along a particular acoustic dimension (e.g., VOT) whereas learners tended to infer one category when given a unimodal distribution of sounds. Parallel results were found with adult learners as well (Maye & Gerken 2000).

The viewpoint that phonetic categorization occurs before word learning was implicitly assumed in early research. However, later research has revealed a considerable temporal overlap between sound and word learning processes during development. For example, the capability of segmenting words in continuous speech has been found among infants as early as six months (Bortfeld, Morgan, Golinkoff, & Rathbun 2005), and this ability continues to develop over the next several months (Jusczyk & Aslin 1995). Word segmentation requires infants to map words heard in isolation onto words heard in fluent

sentences. Previous research has shown that even before a complete set of phonetic categories has been established, infants have already started word segmentation with whatever phonetic categories they have already formed (Jusczyk, 1993a). This raises the possibility that knowledge at the word level may influence speech sound acquisition. Feldman et al. (2011) showed that adult learners increased their sensitivity to distinct vowel categories after being trained with vowels embedded in distinct lexical items (e.g., *gu<u>tah</u>* vs. *li<u>taw</u>*, underlined vowels are the target vowels), whereas adult learners had no sensitivity change after being trained with vowels embedded in the same lexical items (e.g., *gu<u>tah</u>* vs. *gu<u>taw</u>*). Even though this is counterintuitive because the latter condition is actually a minimal pair condition, the authors highlighted the large acoustic overlap between the two vowels and claimed that incorporating word level information may help categorize overlapping categories successfully. The results of the better vowel discrimination generated by non-minimal word pairs support the interactive learning theory proposed by Feldman et al. (2009) that distinct word forms made learners bias towards distinct phonetic categories.

In a pilot study, we replicated the study of Feldman et al. (2011) and examined whether naïve listeners' sensitivity to lexical tones improved when different lexical tones were placed in non-minimal word pairs and minimal word pairs. Tone 2 (T2) and Tone 3 (T3) in Mandarin Chinese were two of the most confusable tones (Shen & Lin 1991, Moore & Jongman 1997). We created a continuum from T2 to T3 in eight steps. Steps 1 to 4 were considered as the category of T2 whereas steps 5 to 8 were considered as the category of T3. In one training condition, we placed the exemplars of T2 and T3 in

minimal word pairs (e.g., *ku1ju2* vs. *ku1ju3*). In the other training condition, we placed the exemplars of T2 and T3 in non-minimal word pairs (e.g., *ku1ju2* vs. *ti1ju3*). Fifteen native English speakers were exposed to the minimal pair disyllables and another 15 native English speakers were exposed to the non-minimal pair disyllables. Each group were exposed to two identical blocks, each of which included 64 target disyllable tokens, half of which had T2 embedded and half of which had T3 embedded. After the first block of familiarization (i.e., listening to the lexical tones passively without doing any task), the participants did an AX tone discrimination task. After the second block of familiarization, the participants did an identical AX tone discrimination task. The result showed that only placing T2 and T3 in the non-minimal word pairs training condition helped improve the sensitivity to T2 and T3 tokens that had larger acoustic similarity (i.e., both T2 and T3 have initial pitch falling followed by final pitch rising); placing them in the minimal word pairs training condition did not. The pilot study's result suggests that the non-minimal pair training condition helps alleviate the acoustic overlap problem for tone categorization as well. The finding in the pilot study suggests that embedding different L2 sound categories in more distinct lexical items or non-minimal word pairs helps adults establish these different sound categories. However, one thing that we need to bear in mind is that the effect of non-minimal pairs was found only when the participants got familiarized with the sound input unconsciously without performing any task. The question is whether such an effect will hold when naïve listeners need to consciously learn the sound categories.

2.4 Previous studies on English speakers' lexical tone categorization

The primary acoustic correlate for Mandarin Chinese tones is F0, i.e., the fundamental frequency of the voice (Abramson 1962, Howie 1976). F0 has been consistently shown to be the dominant cue in adult tone perception (e.g., Klein, Zatorre, Milner, and Zhao 2001, Whalen and Xu 1992). Decades of research on tone perception have shown that its cue-weighting is highly language-specific (Chandrasekaran et al. 2007a, Gandour 1983, Sun and Huang 2012). Previous Multidimensional Scaling (MDS) studies of tone perception have found cross-language differences in dimensions utilized in tone perception (Francis et al. 2008, Gandour 1983). A recent study examined categorical perception of pitch direction, which is defined as a function or curve that tracks the perceived pitch over time, in native and non-native speakers using parametric variation of the direction dimension from level T1 to rising T2 and showed that native speakers exhibited more categorical perception of pitch direction relative to non-native speakers (Xu et al. 2006). Studies examining pre-attentive tonal processing using a neural index of change-detection—the mismatch negativity (MMN)—have demonstrated a superior representation of pitch contour/direction in native speakers of Mandarin Chinese relative to speakers of non-tonal languages (Chandrasekaran et al. 2007b, Kaan et al. 2007). Taken together, a consistent pattern across these studies is that native speakers of Mandarin selectively attend more to pitch contour/direction than non-tone language speakers (e.g., English speakers) whereas non-tone language speakers (e.g., English

speakers) relied more on pitch height, defined as the average pitch value across time, for tone perception than tone language speakers (e.g., native Chinese speakers).

A few studies have already shown English speakers' Mandarin lexical tone identification can be improved by phonetic training. Wang, Spence, Jongman and Sereno (1999) used multi-talker lexical tones to train American English speakers to identify four Mandarin lexical tones, subjects were asked to attend to and identify the pitch patterns without using them to contrast word meaning. In their training, explicit feedback was provided to the listeners after they labeled the tone. With eight sessions of training, each lasting about 40 mins, the identification accuracy increased by an average of 21%. Since the participants in this study had already learned Mandarin Chinese for about one semester, it is unknown how well the training paradigm can help real beginning learners improve their tone identification. Wong and Perrachione (2007) investigated the learning of non-native suprasegmental patterns, three Mandarin Chinese tones in their case, for word identification. Native English-speaking adults without any tone language experience learned to use Mandarin Chinese tones to identify a vocabulary of six English pseudo-syllables superimposed with three pitch patterns (18 words). Successful learning of the vocabulary necessarily entailed learning to use pitch patterns in words. In their study, there were six blocks in the training session and in each block the subjects needed to learn three novel words. Each word was a CVC syllable superimposed with one of three Mandarin Chinese tones (T1, T2 and T4) resynthesized from a high level tone produced by a single male speaker. The syllables used for the stimuli were all legal English syllables because the authors argued that familiar segmental units may ease for

tone learning. After each training block the subjects was quizzed with a word identification task. Upon hearing a sound, the subjects needed to choose the picture that is associated with the sound. They found that the word learning training paradigm can significantly improve tone perception (an increase of tone identification accuracy by 50%) and the individual variance of tone perception after the training can be explained by the fact that lexical learning attainment is mediated by the basic auditory sensitivity to non-lexical pitch patterns as they found that the pre-training non-lexical tone identification accuracy rate was a significant predictor for the final word learning attainment. The key difference between the two training paradigms mentioned above for tone perception is that Wang et al. (1999) focused on improving the low level auditory processing of different lexical tones whereas Wong and Perrachione (2007) emphasized the integration of lexical processing for better tone perception. However, none of these studies examined whether the cue-weighting change occurred after the training on tone perception. The nature of the benefit of high variability for tone categorization was still not very well understood.

Following Wong and Perrachione (2007), Chandrasekaran et al. (2010) implemented the same word learning paradigm to train English speakers' tone perception of all four Mandarin Chinese tones in order to examine whether cue-weighting on pitch direction changed after the training. Same as the training stimuli used in Wong and Perrachione (2007), the pitch tracks were superimposed on six CVC syllables, however, the pitch tracks were not resynthesized but were produced by two males and two females. An INDSCAL analysis, a statistical analysis that infers individuals' psychometric

configuration for perception on one or more dimensions (more details about INDSCAL is provided in Chapter Three), showed that the two dimensions that English speakers used for tone discrimination were interpreted as pitch height and pitch direction when mapped onto the acoustic space of the native Chinese speakers' production. English speakers showed a significant cue-weighting increase on pitch direction after the training. The study demonstrated that the English speakers assigned more weight to the pitch direction dimension for tone perception after learning words with distinct pitch patterns produced by four different talkers. The length of the training was comparable to Wang et al.'s (1999) study: 9 sessions over two weeks with each session lasting about 30 mins. One thing worth mentioning is that, so far in most studies, the native Chinese speakers and non-native Chinese speakers' cue-weighting for tone perception is derived from the RT for the discrimination of synthesized tones rather than naturally produced tones (Gandour 1983; Wong and Perrachione 2007; Chandrasekaran et al. 2010). In the naturally produced Mandarin tones, low dipping tone (T3) and high falling tone (T4) are often accompanied with creakiness and also different tones differ in terms of duration (Keating and Esposito 2006, Moore and Jongman 1997, Sereno et al. 2011). Thus, in order to control the voice quality and duration factors, most studies on tone perception used synthesized tones. In terms of tone learning, when multi-talker stimuli are used as training stimuli, there are voice quality differences among different tones. It is possible that the learners use the creakiness of certain tones as the cues for tone categorization during the training. With such additional acoustic cues (i.e., creakiness) in the training stimuli, they may decrease the importance of pitch height to a certain degree and allows

more attention to be shifted towards tone pairs that differ in pitch direction, allowing native English speakers to shift cue-weighting towards pitch direction after multi-talker training.

2.5 Outline of the current study

The recent psycholinguistic findings showed that: (a) a multi-modal phonetic training paradigm that encodes visual, interactive information is more effective in training L2 learners' perception of novel categories, (b) decreasing the acoustic variance of a phonetic dimension allows the learners to more effectively shift the perceptual weight towards this dimension, (c) using an implicit word learning task in which the words are contrasted with different lexical tones improves naïve listeners' categorization of Mandarin Chinese tones, and (d) getting familiarized with novel sound categories embedded in non-minimal pairs improves the sound discrimination more than embedding the sound categories in minimal pairs under the condition that the participants do not need to perform any task during the familiarization phase.

Based on all these psycholinguistic findings, in our current study, we used a video game training paradigm to train American English speakers' categorization of the four Mandarin Chinese tones in either minimal pairs or non-minimal pairs by using two types of training stimuli, namely, variance manipulated training stimuli and multi-talker training stimuli. For our study, we tested whether the unequal variances can change naïve listeners' cue-weighting for the perception of Mandarin Chinese tones. Based on previous behavioral and neurolinguistic studies on American English speakers' tone perception,

we expect English speakers to assign more weight to pitch height than pitch direction before the training. We then provide them with a training dataset in which the four Chinese Mandarin tones have a small variance on the pitch direction dimension and a large variance on the pitch height dimension. In addition, there was no overlap on pitch direction among the four tones and there was a significant overlap on pitch height among the four tones. If the decrease of within-category variance on a certain acoustic dimension can elicit the increase of weight on that particular acoustic dimension, then we would expect the weight to shift from the pitch height dimension to the pitch direction dimension after training, which will be more nativelike in terms of tone perception. Previous studies showed that the pitch height (averaged pitch value in Hz) was spread wider than pitch direction (pitch slope calculated as the pitch change rate in Hz/s) within each Mandarin tone in a spontaneous speech corpus consisted of ten male and ten female speakers' tone production (Coster and Kratochvil 1984). Therefore, we only resynthesized the tone stimuli in a way that mimics the real distribution of four Mandarin tones in the space of pitch height and pitch direction, namely, tones are spread wider on pitch height than pitch direction within each lexical tone. Another reason we did not manipulate the variance in other ways such as making the variances on pitch height and pitch direction equal or the variance on pitch direction larger than the one on pitch height is that the ultimate goal is to allow native English speakers to learn phonetic sound categories (in this case Mandarin lexical tones) rather than non-phonetic or non-speech categories. In other words, with the purpose of phonetic training, we need to make the training tone stimuli similar to the real lexical tones. That is why we resynthesized the

training tone stimuli with naturally produced tones. It at least guaranteed the naturalness of the pitch tracks of the tones.

In the current study, we also examined the efficiency of the video game training paradigm to see whether a shorter period of training can reach a comparable amount of improvement in terms of tone categorization as found in the previous studies. In our study, each participant received a total of two hours training in four different days. We also embedded different lexical tones in either minimal word pairs or non-minimal word pairs in order to test whether more distinct lexical items can help English speakers establish four different tone categories. In addition to integrating variance manipulated stimuli with the video game training paradigm, we also integrated multi-talker stimuli with the video-game training paradigm to investigate whether the combination of high variability training and video-game training can efficiently improve native English speakers' tone categorization and shift their cue-weighting towards pitch direction after the training. We also aimed to examine whether more nativelike cue-weighting was correlated to the overall tone discrimination performance and the discrimination of certain tone pairs (e.g., T1 differs from T3 mainly in terms of pitch height; T2 differs T4 mainly in terms of pitch direction). Finally, we examined which training condition allows better generalization of tone identification to new talkers.

2.6 Research questions and Hypotheses

The current study addresses three themes: (1) the efficiency of the multi-modal phonetic training paradigm for the training of L2 tone categorization; (2) the effects of

talker variability, variance, minimal pairs and non-minimal pairs for shifting cue-weighting for tone perception; (3) the effectiveness of different types of training stimuli on L2 tone categorization (e.g., tone identification of both old and new talkers; tone discrimination).

The specific research questions are listed below:

(1) Does videogame training with only a total of two hours allow naïve listeners to learn four Mandarin tones without any explicit feedback? In other words, can naïve listeners discriminate and identify the four tones reasonably well after the training?

(2) Does a smaller variance on pitch direction and a larger variance on pitch height help naïve listeners shift their cue-weighting towards the pitch direction dimension after the training?

(3) Does multi-talker training help naïve listeners shift their cue-weighting towards the pitch direction dimension after the training?

(4) What types of training stimuli (or training conditions) produce the best tone discrimination result? What types of training stimuli produce the best tone identification result?

(5) What training conditions allow naïve listeners to generalize their tone identification to new talkers?

(6) Will the learners whose cue-weighting for tone perception becomes more nativelike perform better on the tone categorization task than the

learners whose cue-weighting for tone perception are still less nativelike?

(7) What is the relation between the video game performance and the ultimate tone categorization performance?

2.7 Hypotheses

We predict that the multi-modal phonetic training paradigm, being an object-oriented task, helps learners improve L2 sound categorization more efficiently as it better captures learners' attention. In terms of the potential for triggering a cue-weighting shift for tone perception, we predict that using multi-talkers' tone tokens as training stimuli will make naïve listeners shift their cue-weighting towards pitch direction based on the finding in Chandrasekaran et al. (2010). In terms of the effect of variance on perceptual cue-weighting, we predict that the variance effect for cue-weighting shift at the segmental level can be generalized to the suprasegmental level. Based on the finding of the effectiveness of non-minimal word pairs for improving sound discrimination, we predict that embedding contrastive tones in non-minimal word pairs may help improve tone discrimination more than embedding the tones in minimal word pairs with some caution because the effectiveness of the non-minimal pairs found in an unsupervised learning may not be generalized to a semi-supervised learning condition such as the videogame used in the current study. Finally, we predict that a good video game player will be a good learner in the setting of our current phonetic training paradigm.

CHAPTER THREE: METHODS AND EXPERIMENTAL DESIGN

3.1 An animal feeding video game

We created a video game in a 2D space. In the game, the participants needed to select the correct food to feed four different animals. There were four animals—1) cat; 2) monkey; 3) dog and 4) rabbit. The animals' favorite foods were shown at the top of the screen—1) fish; 2) banana; 3) bone and 4) carrot. The animal appeared one at a time and ran across the computer screen. Each animal was associated with a specific lexical tone: cat—T1; monkey—T2; dog—T3; rabbit—T4. For each lexical tone, there were 72 exemplars/tokens. During each trial, an exemplar of a lexical tone was randomly selected and played repeatedly to the participant together with the appearance of the animal. At the beginning, the animals were clearly visible, as shown in Fig.1. As the game progressed, it became more and more difficult to identify the animal visually, as shown in Fig.2. To make it difficult to identify the animals visually, we only showed part of the animal (e.g., only the head) in a vehicle. We used 7 different speed levels. The higher the game level, the faster the animal moves across the screen. From game level 7 to level 10, however, the speed did not change as the animals were completely invisible starting from level 7. The lexical tone information was available auditorily throughout the game. In other words, at the beginning stage of the game, players can simply depend on the visual information to feed the animal the correct food; for the later levels of the game, however,

the players needed to rely more on the auditory information to identify the animal and feed it.

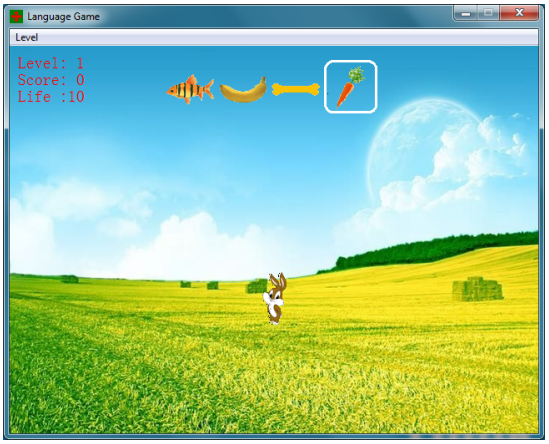

Figure 1. Video game at level 1

Figure 2. Video game at level 5

There were 10 levels in total. Each level required 720 points to clear. 10 points were added to "Score" when an animal was fed the correct food; meanwhile one point was added to "Life". One point was deducted from "Life" when the animal was fed the wrong food, and "Score" did not change. If the animal ran through the screen without being fed, one point was deducted from "Life" as well. When "Life" reached zero, the game was over. The purpose of "Life" was to allow the participants to provide the wrong response for a certain number of times. It may allow the participants to track the mistakes they made. Life was initialized with 10 points. For every level of the game, each animal appeared 18 times, thus, all four animals appeared 72 times in a cycle. During this cycle, if the participant did not reach the required score (720 points) to pass the level, another

cycle within the same level would begin. The order of the animals' appearance was randomized.

In terms of the operation, participants needed to use their left hand to press keys 1, 2, 3 and 4 to choose the food. The selected food was highlighted. Then the participants needed to use their right hand to control the mouse to aim at the animal and left click to feed it.

In order to track the individual differences when playing of the game, each participant's correct and incorrect responses were recorded for each level of the game. In the training period, each subject played the video game for 4 sessions, each of which lasted 30 mins except the last session, which only lasted 15 mins because the participants needed to do the post training tone discrimination and identification tasks. This video game training paradigm was an implicit learning of lexical tone in nature as no explicit feedback or information about the tones was provided to listeners during the game. In order to play the game well, the naive listeners had to draw their attention to the sounds with distinct tone categories. The instruction given in the video-game training was as follows:

> "***Goal***
>
> *There are four animals—a cat, a monkey, a dog or a rabbit — that will appear on the screen. They are running from the left of the screen to the right. Your task is to select each animal's favorite food to feed the animal. The foods you can select include: a fish for the cat, a banana for the monkey, a bone for the dog or a carrot for the rabbit.*
>
> *The game has 10 levels in total. Every time you feed the animal the correct food, the score will increase by 10 points. If you either feed the animal with wrong food or let it run across the screen without being fed, the score will stay the same but your life will be reduced by one*

point and by more points later in the game. If your life is reduced to zero, you need to choose a level at which to replay the game.

Operation

- *Put on the headphones. There will be sounds playing during the game.*
- *Use the left hand to press keys 1, 2, 3 and 4 to select the food. The selected food will be highlighted.*
- *Use the right hand to move the mouse to aim at the animal and left click to feed it.*

A tip for playing the game

The game will get more and more challenging. You need to develop a strategy by using whatever information available in the game in order to clear the difficult levels. If game over, choose a level you are comfortable with to restart playing the game."

3.2 Experiments

Using this video game, we conducted five experiments that used different types of training stimuli for native English speakers to learn L2 Chinese tone categories. The first set of experiments used monosyllables as training stimuli whereas the second set of experiments used disyllables as training stimuli. The participants did two AX discrimination tasks (one used monosyllables as test stimuli, one used disyllables as test stimuli) before and after the training. The pre- and posttests served the purpose of examining the participants' sensitivity to lexical tones in different syllable contexts. Also, the first AX discrimination task served the purpose of examining participants' cue-weighting on pitch height and pitch direction. The participants also did a word identification task after the training. This task served the purpose of examining whether

the participants were able to associate the novel sounds to the animals used in the training. A group of native Chinese speakers also did the experiment as a native control group.

3.2.1 Participants

For naïve listeners, there were one control condition and four training conditions and 10 participants were recruited and tested for each condition. The information about the participants in terms of age, gender and length of formal music training is illustrated in the following table.

Table 2. The age range, number of males and females and the number of participants who had formal music training more than six years, less than six year and no music training.

Experiment	Age (mean and range)	Male	Female	Formal music training over six years	Formal music training under six years	No music training
Control	21, 18-23	3	7	2	6	2
Variance-manipulated	24, 18-29	5	5	0	8	2
Multi-talker	25, 20-32	5	5	0	7	3
Minimal pair (disyllable)	24, 19-27	4	6	0	7	3
Non-minimal pair (disyllable)	22, 18-25	4	6	0	8	2

None of the participants learned any tone language before. We also tried to recruit participants who had formal music training as little as possible. Previous research that studied non-tone language speakers' perception of Mandarin tones excluded participants who had formal music training over six years because long musical training may increase tone discrimination abilities (Chandrasekaran et al. 2010). All except two participants in our study had less than six years of formal music training; both were in the control group.

3.2.2 Experiment 1a—Non-native control

Experiment 1a aimed to establish a baseline for naive listeners' tone categorization. As training stimuli, we used four monosyllables /sa/, /fa/, /ma/ and /na/ recorded by a male native English speaker. All the segments in the CV syllables exist in English. Thus, we expect native English speakers to use the segmental information to play the video game. We used PSOLA (pitch-synchronous overlap-and-add) in Praat to normalize the pitch of the four monosyllables, making them all have a high level tone. We also manipulated the duration of the initial consonants proportionally to its original length, using the lengthen function in Praat, in order to add some variability in the training tokens. The vowel duration of the tokens were normalized to be 300ms. Each monosyllable had 4 tokens. In total, there were 16 monosyllable tokens without lexical tones.

3.2.3 Experiment 1b—Variance manipulation training

Experiment 1b aimed to test the robustness of variance manipulated stimuli training in tone categorization and cue-weighting shift.

The stimuli consisted of four lexical tones that had a smaller variance on pitch direction relative to the variance on pitch height. Each tone had 18 exemplars. In total, there were 72 tokens. The pitch direction was quantified as (pitch offset-pitch onset)/duration whereas the pitch height was quantified as the f0 value averaged across 11 time normalized pitch values using Yi Xu's TimeNormalize Praat script (Xu 1997). To make the variance manipulated monosyllables with different lexical tones, we first had a male speaker who had a middle-range fundamental frequency record the four lexical tones on a monosyllable 'yu' /y/ (a high front rounded vowel) in citation form. The reason we used /y/, which does not exist in English, is to avoid the effects of word frequency, neighborhood density and other lexically related factors. We selected three tokens for each lexical tone with a slight pitch direction difference from the recorded tokens. We made sure no tones had any overlap in terms of pitch direction. Then we used PSOLA in Praat to shift the pitch tracks of each lexical tone so that six different pitch height values for each lexical tone were derived. Finally, we extracted the pitch tracks and used a single 'yu' token to resynthesize the four tones so that all acoustic cues were controlled except tones. The duration of the vowel was normalized to be 300ms. The pitch tracks of four lexical tones in the base tokens are illustrated in the following graph. The three pitch tracks for each lexical tone were produced by three different male native

Chinese speakers.

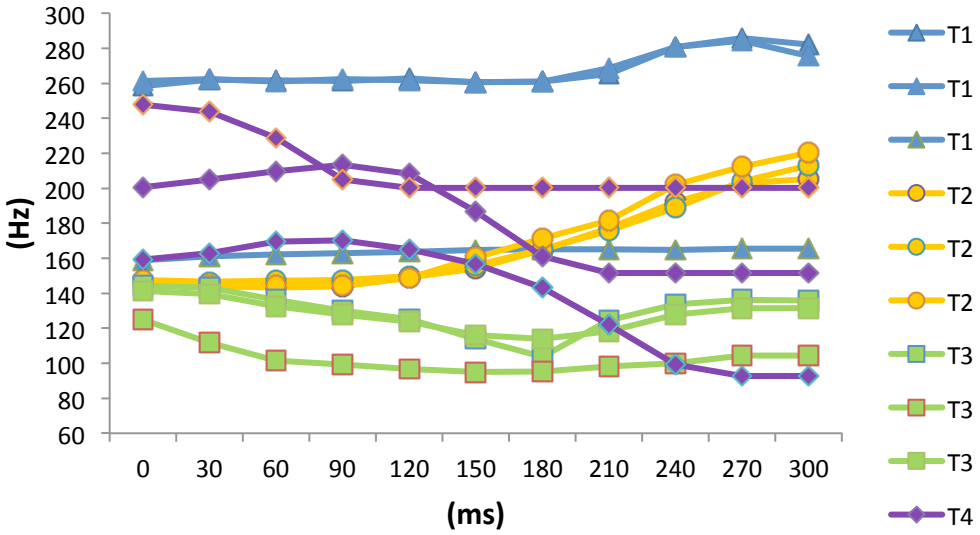

Figure 3. Pitch tracks of four different lexical tones in the base tokens for resynthesizing variance manipulated training stimuli.

Qualitatively, the pitch height of the four lexical tones overlapped substantially but the pitch direction of the four lexical tones were quite distinct. After creating the variance manipulated tokens with different lexical tones, five native speakers of Mandarin Chinese listened to the resynthesized stimuli and did a tone labeling task. For an exemplar to be used as a stimulus, all five native Mandarin Chinese speakers needed to identify its tone correctly. One thing that needs to be pointed out is that the T3 we used for the training stimuli was a low dipping tone. Following Chandrasekaran et al. (2010), we simplified the tone direction calculation of T3 by subtracting the pitch offset from the pitch onset divided by the vowel duration. The pitch height and pitch direction values for the four lexical tones are summarized in Table 3 and Table 4, respectively. Fig. 4 shows

the distribution of the 18 exemplars for each lexical tone (3 pitch direction x 6 pitch height). The three base pitch tracks were produced by three male native Chinese speakers respectively. For each pitch track, we shifted the pitch tracks upward or downward with an equal step of 10 Hz, making six different pitch height values for each base pitch track. One thing worth mentioning is that pitch height is correlated to the identification of gender in real speech. In order to tease apart the effect of talker variability and variances on different acoustic dimensions, we made the range of the pitch height of the resynthesized tone stimuli within the pitch range of the male voice. We made sure that all resynthesized stimuli sound like male voices to the listeners.

Table 3. Six pitch height values in Hz (average f0) and the standard deviation of pitch height in jnd for four lexical tones.

	1	2	3	4	5	6	Standard deviation (jnd)
T1	210	200	190	180	170	160	18.7
T2	190	180	170	160	150	140	18.7
T3	160	150	140	130	120	110	18.7
T4	200	190	180	170	160	150	18.7

Table 4. Three pitch direction values in Hz/s (pitch slopes) for four lexical tones

	1	2	3
T1	14	47	79
T2	145	202	268
T3	-22	-65	-103
T4	-240	-354	-441

Table 5. Quotients of rate of change converted from Hz/s among three tokens for each lexical tone

	Quotient between token 1 and 2	Quotient between token 1 and 3	Quotient between token 2 and 3	Standard deviation (jnd)
T1	3.36	5.64	1.68	1.9
T2	1.39	1.85	1.33	0.3
T3	2.95	4.68	1.58	1.6
T4	1.48	1.84	1.27	0.3

Since the units of pitch height and pitch direction are different (pitch height: Hz; pitch direction: Hz/s), we could not directly compare the variance of pitch height with that of pitch direction. We can only compare the variances on the acoustic dimensions indirectly. One way to make the comparison is to follow Smits et al. (2006) and Goudbeek et al. (2005) where the variances on duration and formant were quantified in terms of just noticeable difference units converted from different scales (Equivalent Rectangular Bandwidth (ERB) for formant and psychological duration D for duration, both of which are logarithm transformations). In the research of jnd for pitch, previous studies showed that listeners are extremely good at distinguishing successively presented level pure tones that differ in frequency. For example, Harris (1952) showed that it was not uncommon for the frequency differential limens of pure tones to be less than 1Hz. Flanagan and Saslow (1958), using synthetic vowels in the frequency range of a male speaker, reported the differential limen to be between 0.3-0.5Hz, and this result was replicated by Klatt (1973). The jnd may be slightly higher for differentiating non-level tones (Klatt 1973). Based on these studies, we used 1 Hz as the jnd for differentiating the

same lexical tones but with different pitch heights. Then the standard deviation of the pitch height for the four tones is 18.7 jnds, as shown in Table 3.

As for jnd for distinguishing pitch contours, it is usually calculated in terms of the quotient of two pitch contours' rate of pitch changes. Pollack (1968), using synthesized falling tones with constant initial frequencies of 125-1000 Hz and durations of 0.3 to 4 seconds, reported differential thresholds of two pitch changes from 0.5 to 4 seconds in terms of the quotient of the their rates of change in Hz/s. He showed that the minimum quotient was around 2 for the stimuli. Klatt (1973) studied the differential thresholds of pitch changes in speech-like signals and reported that listeners could distinguish a 135Hz to 105Hz f0 fall from a 139Hz to 101Hz f0 fall, both with a 250ms duration. The differential threshold here, if converted to the quotient of rates of change (1.27), was even better than the results in Pollack (1968). Since the duration of the resynthesized stimuli in the current study was 300 ms, we used the quotient of rates of change (1.27) as the jnd for differentiating the same lexical tone but with slightly different pitch direction. We first calculated the ratio between the slopes of each token pair within each lexical tone, thus converting Hz/s to quotient. The quotient values for each lexical tone are shown in Table 5. As we can see, the quotient between each different tokens within each lexical tone is larger than 1.27. Thus, theoretically the differences between different tokens of the same lexical tone are perceivable. The standard deviation on pitch direction for each lexical tone became very small, as shown in Table 5. In terms of jnd, among the training stimuli in variance-manipulation training, pitch height indeed has a greater variance than pitch direction.

In terms of integrating the variance manipulated stimuli with the video game, after creating the exemplars of each lexical tone, they were placed in four different sound folders, each of which corresponded to an animal. When an animal appeared in the game, the exemplars of a particular lexical tone were randomly selected from the corresponding sound folder without replacement and played repeatedly.

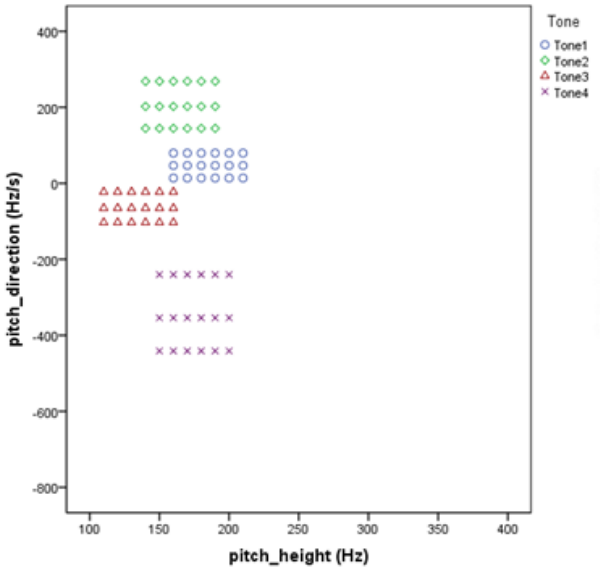

Figure 4. Distribution of variance manipulated exemplars of four lexical tones.

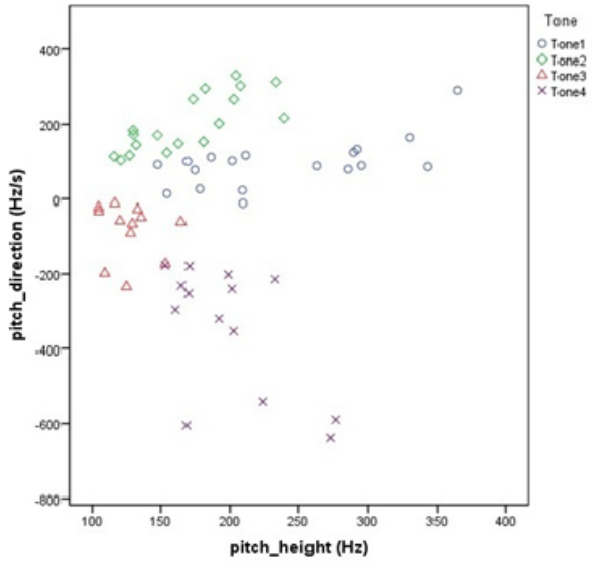

Figure 5. Distribution of multi-talker exemplars of four lexical tones.

3.2.4 Experiment 1c—Multi-talker training

Experiment 1c aimed to test the robustness of multi-talker stimuli training in terms of tone categorization and cue-weighting shift.

The stimuli that included 18 exemplars of each lexical tone on the monosyllable 'yu' were produced by 9 different native speakers of Mandarin Chinese (5 males and 4 females). In total, there were 72 tokens. Fig. 5 shows the exemplars of each lexical tone

we had recorded in the pitch height and pitch direction acoustic space. We can see that naturally produced lexical tones still had substantial overlap on pitch height but less overlap on pitch direction. Thus, we expect participants' cue-weighting to shift towards pitch direction after the multi-talker training. Comparing the variances of pitch height and pitch direction in Fig. 4 with those in Fig. 5, we can see that the naturally produced lexical tones are less distinct on pitch direction relative to the variance manipulated lexical tones. The indexical information included in the multi-talker training stimuli such as gender and voice quality (e.g., creakiness vs. non-creakiness) also contributes to the characteristics of the variances on pitch height and pitch direction. For example, males on average have lower f0 than females. These f0 differences make the training stimuli more spread out on pitch height relative to pitch direction. Although this makes the tone categories more variable on pitch height, the gender difference is easy to detect based on pitch height, voice quality and possibly other cues. Once the gender is identified for each training token, it means that the highly variable tone stimuli are normalized. It is likely that the participants make use of pitch height for the normalization process, after which they realize pitch direction is the dimension that is more reliable for differentiating tone categories. In other words, pitch height may be used for non-phonetic purposes such as gender identification to a certain extent if not completely, whereas pitch direction is used more for the purpose of phonetic categorization. In order to make the participants use f0 to the maximum degree, we normalized the amplitude and duration of the multi-talker tone tokens (amplitude: 70 dB; duration: 300ms). The normalized amplitude and duration were the same as the ones of the variance manipulated training stimuli in Exp. 1b.

3.2.5 Experiment 1d—Native control

Experiment 1d aimed to examine native speakers' cue-weighting for lexical tone perception. Ten native speakers of Mandarin Chinese were recruited. However, they did not participate in any training. They did a tone discrimination task and we used INDSCAL analysis to calculate their cue-weighting for tone perception, using monosyllables as the test stimuli (see more details in Chapters Four).

3.2.6 Experiment 2a—disyllable minimal pair training

Experiment 2a aimed to test whether embedding variance manipulated stimuli from Exp. 1b (variance manipulation training) in minimal word pairs can help improve tone categorization. We embedded the variance manipulated monosyllable 'yu' with different tones in minimal pair disyllables—*ta1yu1, ta1yu2, ta1yu3 and ta1yu4*. The duration of *ta1* was 250ms. Each disyllable had 18 exemplars (1 *ta1* x 18 *yu*). In total, there were 72 disyllable stimuli. For the disyllables, we concatenated the monosyllable *ta1* recorded by the same speaker who recorded the monosyllable *yu*. We adjusted the offset of the pitch of *ta1* to make it closer to the onset of the following variance-manipulated lexical tones in order to mimic a more natural pitch transition from T1 to the different following lexical tones based on the naturally produced tonal transition in disyllables. One thing worth mentioning is that we did not make any changes to the variance-manipulated tone tokens on the second syllable, as we want to control the effect of variance-manipulation and examine whether the context of disyllable minimal pairs

can further improve naïve listeners' tone categorization. Therefore, the pitch transition from the first syllable to the second syllable still may not be entirely natural.

3.2.7 Experiment 2b—Disyllable non-minimal pair training

In Experiment 2b, we again used variance manipulated *yu* with different tones as the training stimuli. In Experiment 2b, we concatenated the variance manipulated *yu* with monosyllables to make four non-minimal word pairs—*ta1yu1, ku1yu2, po1yu3, ti1yu4*, which differed not only in tones on the syllable *yu* but also in segments of the first syllables. We selected four CV syllables that differ both in terms of the consonants and vowels as the preceding syllables in order to maximize the differences among the four disyllables. We had the same male speaker who recorded the monosyllable *yu* with different lexical tones record the preceding syllables *ta1, ku1, po1 and ti1*. We shifted the pitch of the preceding T1 to make a more natural pitch transition to the second syllable. For each lexical tone on 'yu', it had 18 disyllable tokens (e.g., 1 *ta1* x 18 *yu1*, 1 *ku1* x 18 *yu2*, 1 *po1* x 18 *yu3* and 1 *ti1* x 18 *yu4*.). In total, there were 72 non-minimal pair disyllable tokens. According to Feldman et al. (2011) and our own pilot results, we expect that the participants' tone categorization performance, especially tone discrimination performance, to significantly improve after the training, and even more so than the minimal word pair training in Experiment 2a, due to their ability to implicitly track the sound category distribution in the non-minimal-pair lexical item. But we cannot rule out the possibility that the participants simply ignore the tone information but only use the first syllable information to play the game. In that case, it is hard to predict

whether the non-minimal pair training condition can produce better tone categorization results than the minimal pair training condition.

3.3 Pretest and posttest

Two AX discrimination tasks were used in the pretest and posttest for all six experiments (Native Chinese speakers in the native control condition only participated in the pretest AX discrimination tasks but did not participate in the videogame training). In addition, a word identification task was used after the video-game training for the five native English speaker groups but not for the native control group.

3.3.1 Discrimination task for cue-weighting calculation

The first task was a speeded AX discrimination task that served the purpose of evaluating naive listeners' cue-weighting on pitch direction and pitch height dimensions. The discrimination task consisted of 6 blocks, each of which contained 24 stimulus pairs. The 24 pairs in each block include 12 different tone pairs ($n(n\text{-}1)=4x(4\text{-}1)=12$; e.g., T1 vs. T2, T1 vs. T3, T1 vs. T4, etc.) and 12 identical pairs (4 identical tone pairs are repeated three times). Thus, all participants listened to $24 \times 6 = 144$ pairs of the form /i-i/ whose tones are either the same or different. Here, we use the monosyllable 'yi' /i/ for the test, which is different from 'yu' /y/ used in training. Although there were 144 stimulus pairs, only 4 tone tokens are used as the stimuli. For this discrimination task, a resynthesized version of four lexical tones modeled on the basis the previous acoustic study of Mandarin Chinese tones (Xu 1997) were superimposed on an /i/ token produced

by a male native speaker of Mandarin Chinese, using the PSOLA method implemented in Praat. The duration of all four tone tokens was 250 ms. Following Chandrasekaran et al. (2007b), the parameters of the four lexical tones are summarized in Table 6 and the pitch tracks of the four lexical tones were plotted in Fig. 6.

Table 6. Acoustic characteristics of tone tokens used for speeded AX discrimination task.

	F0 parameters						
Tone	Onset	Offset	Slope[a] (Offset-Onset)	Slope[b] (Onset-TP)	Slope[c] (TP-Offset)	TP (ms)	Average
T1	129	128	0.00	0.02	-0.02	125	129
T2	109	136	0.11	-0.04	0.17	71	117
T3	104	109	-0.02	-0.13	0.19	133	96
T4	140	90	-0.20	0.08	-0.27	144	124

Note. Onset, offset, and average F0 values are expressed in Hertz (Hz). All slope values are expressed in Hz/ms. T1, T2, and T3 refer to the Mandarin high level, high rising, and low dipping tones, respectively. TP, expressed in milliseconds, refers to turning point, i.e., time at which the contour changed direction. F0 = voice fundamental frequency; Δ F0 = change in Hz from onset to turning point.

a. Overall slope, measured from pitch onset to offset.
b. Slope from the onset to TP. Since the level tone T1 has no clear turning point, slope was measured from onset to 125 ms (50% duration). Both T3 and T4 have negative slopes (i.e., falling F0 contour). T2 has a positive slope (i.e., rising F0 contour).
c. Slope from the TP to offset. Since the level tone T1 has no clear turning point, slope was measured from 125 ms (50% duration) to offset. Both T2 and T3 have positive slopes (i.e., rising F0 contour). T4 has a negative slope (i.e., falling F0 contour).

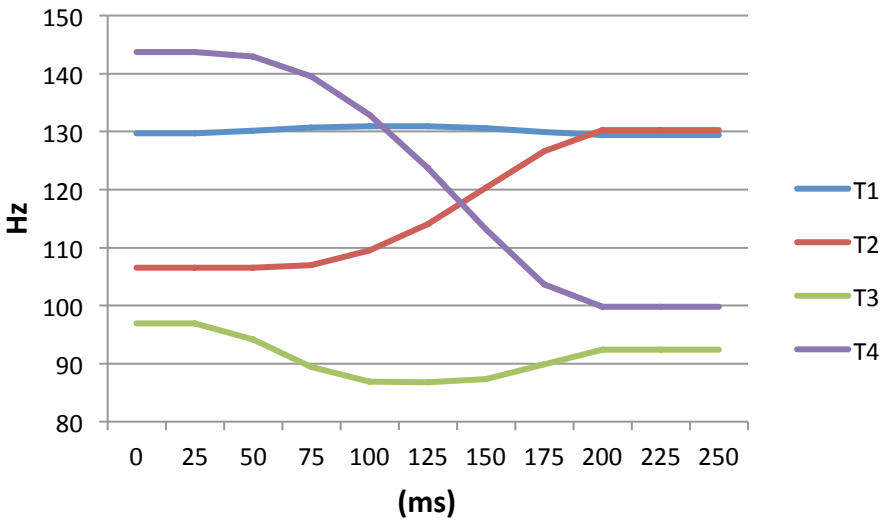

Figure 6. Pitch tracks of four lexical tones used for the speeded AX discrimination task.

In the speeded discrimination task, the participants were asked to judge whether the stimulus pairs are the same or different as fast as possible. In order to balance the response speed and the attention paid to the task, we provided feedback to the participants after each trial to inform them whether their response was correct or incorrect. Accuracy and reaction time (RT) were recorded and used for the sensitivity d' (Macmillan and Creelman 1991) and INDSCAL analyses (Carroll and Chang, 1970), respectively. The d' analysis was used to examine participants' sensitivity to lexical tones in monosyllables using hit and false alarm ratios. The INDSCAL analysis was used to explore the cue-weighting on two expected dimensions, namely, pitch direction and pitch height for each individual participant (More details about INDSCAL are provided in Chapter Four).

3.3.2 Discrimination task for examining sensitivity to lexical tones in disyllables

The second task is a non-speeded AX discrimination task that used disyllables as test stimuli. It aimed to examine how well English speakers can discriminate the target syllable *yu* (/y/) with different tones on the second syllable of a disyllabic word before and after training. For the disyllable AX discrimination task, the subjects were asked to judge whether the second syllables in each disyllable pair were the same or different. The tone discrimination in disyllable context was deemed to be more difficult than the tone discrimination in monosyllable context because each lexical tone had variability when preceded by different tones. In order to avoid discouraging the participants and control the time of the experiment, we did not provide feedback after each trial. If English speakers can judge the same lexical tone category to be the same regardless of the slight physical differences between the tone tokens, then it may suggest that the participants have formed tone categories. In the discrimination task, a set of disyllables *ma* (with T1, T2 and T4) followed by *yu* (with T1, T2, T3 and T4) were used as test stimuli (*ma* with T3 was excluded due to tone sandhi with the following T3). There were therefore 12 different disyllables. Two male native speakers of Mandarin Chinese recorded all 12 different disyllables in citation form. The amplitude of the disyllables were normalized to be 70 dB. The duration of the target syllable *yu* was normalized to be 450 ms[1]. In total,

[1] All disyllables used as the test stimuli were recorded in citation forms. Thus, there was final lengthening, which caused the target syllable 'yu' to be relatively long. We originally shortened it to be 250ms so that the length was the same as the duration of the monosyllable test stimuli used in the first AX discrimination task. However, the shortened disyllables sounded like a word spoken with a fast speaking rate. Thus, we lengthened the target syllables' duration to 450ms so that the stimuli sounded like a word spoken with normal speaking rate.

there were 144 disyllable pairs. One stimulus in the pair was produced by one male speaker and the other stimulus in the pair was produced by another speaker. In other words, each tone pair was produced by two different male speakers.[2] Among these 144 disyllable pairs, 36 pairs had identical tones on the second syllable and 108 pairs had different tones on the second syllable. We presented the 36 pairs with identical tones on the second syllable three times so that there were 108 stimulus pairs with identical tones on the second syllable and 108 stimulus pairs with different tones on the second syllable. In total, 216 disyllable pairs were used. We used hit and false alarm ratios to calculate the d' score for each participant.

3.3.3 Word identification task

After training, a word identification task was used to test how well the English speakers learn the four words after the training. Participants were expected to find that the sounds played in the game were associated with four different animals, the participants should know that those sounds accompanying the animals were not meaningless. The word identification task served the purpose of examining whether implicit word learning led to sound to meaning association. The word identification task consisted of two identical blocks. Each block contained 16 tokens (4 words x 4 repetitions =16). The first block used 16 word tokens used in the training. The second

[2] The reason we used two male speakers' voices was because in a pilot study, we found that the English speakers' tone discrimination in disyllables was relatively high when the tone pairs were produced by a single male speaker. The high accuracy of the discrimination task in the pretest may disguise the training effect. Thus, we used two male voices for the discrimination task in order to make the task more difficult.

block used 16 word tokens produced by another male and a female speaker. On hearing a sound, the subjects were asked to label the sound with an animal. The second block was used to test whether word identification generalizes to new talkers. In total, there were 32 word tokens for the word identification task.

CHAPTER FOUR: DATA ANALYSIS AND RESULTS

In this chapter, we first provide some details about INDSCAL analysis that is used for calculating participants' cue-weightings for tone perception and then we report the results of the experiments. The results are organized as follows: Section 4.2 reports the cue-weighting results of the non-native control group and four trainee groups before and after the video-game training. The cue-weighting result of the native Mandarin Chinese speakers is also reported. Section 4.3 reports different trainee groups' sensitivity to the four lexical tones in monosyllable and disyllable contexts before and after the training. Section 4.4 reports different trainee groups' word/tone identification result after the training. Section 4.5 reports a regression analysis that uses two parameters that measured the video-game performance to predict the ultimate tone discrimination sensitivity d' for disyllable context and tone identification accuracy rate. Section 4.6 reports the relation between the cue-weighting and tone categorization performance.

4.1 Description of INDSCAL procedure

Individual differences multidimensional scaling (INDSCAL) is an extension of general MDS techniques that preserves individual differences. MDS applies a mathematical model that is similar to principal components analysis. Using the proximity data, MDS minimizes a loss function to place the perceived differences between stimuli points in a multidimensional space. The MDS loss function is designed so that the

computed distances between stimuli are as faithful to the actual proximities as possible (Borg & Groenen 2005). This function is computed in an iterative process resulting in a minimized badness-of-fit measure. Once an Euclidean space has been mapped, standard multivariate techniques can be used to determine the number of dimensions contained within the space, place it in a coordinate system, and locate each stimulus with respect to the resulting coordinate axes.

INDSCAL makes one additional assumption. Based on the possibility that individuals (or groups of people) may perceive given stimuli differently, INDSCAL assumes that differences between individuals correspond to differences in the dimensional salience along which stimuli may be classified. That is, individuals are thought to use the same set of dimensions to make their ratings, but to different extents (Arabie, Carroll, & DeSarbo 1987). Classical MDS treats all subjects as equal, eliminating individual differences. INDSCAL preserves these differences by using the individual configurations as its starting point. This results in a complex loss function that first computes a stimulus configuration for each participant and then minimizes the computed distances between stimuli across all participants to produce the most parsimonious group stimulus configuration. This iterative process creates the overall group solution by stretching or shrinking every individual configuration's (also called private perceptual space) axes to match as closely as possible all other individual configurations. Thus, the group solution is computed from a linear combination based on every person's private perceptual space.

One major advantage of INDSCAL is that the weights associated with the individual configurations reflect individual differences, and between-subject comparisons can be made. This additional benefit makes the INDSCAL model ideal for exploring language background related perceptual differences on novel stimuli. These weights are computed in a vector space. After normalizing the vector length, the direction of each subject's weight vector tells us the subject's preference of the perceptual dimension. The deviation of this angle from a 45-degree bisector represents the relative preference of the participant for one dimension more than the other. For example, if one participant prefers using Dimension 1 (Dim 1 henceforth) to discriminate tone stimuli more than Dimension 2 (Dim 2 henceforth), that individual's weight vector would be tilted closer to the first than to the second perceptual dimension. This type of information can be used to answer questions about individual differences regarding the use of one dimension versus another.

Another important reason we use INDSCAL rather than logistic regression to examine individuals' perceptual weights is that the pitch values of a tone change consistently as a function of time. The acoustic characteristics make tone qualitatively different from segments, which have (e.g., consonants and vowels) relatively stable acoustic characteristics. It is very difficult to incorporate any time component in the acoustic dimension to build a linear function to predict values that correspond to different tones or classify different tones by using a non-linear transformation (e.g., a logarithm function). Thus, so far, INDSCAL is the most appropriate analysis for examining cue-weighting for tone perception.

In the current study, for each experiment, twenty (10 participants x 2 pretest, posttest) separate 4 stimulus tones x 4 stimulus tones symmetric data matrices were used as input to the INDSCAL analysis. Each data matrix contained distance between tone pairs estimated based on 1/RT (Shepard 1978, Huang 2001). It means that the longer the RT (Reaction Time) is, the smaller the perceptual distance between the two tones is. The INDSCAL analysis used 1/RT as dependent variable. INDSCAL analyses of these 20 dissimilarity matrices were performed at n where n =1, 2, 3 dimensionalities in order to determine the appropriate number of dimensions underlying the distances among the four tones or objects in a perceptual space. All INDSCAL analyses were made by using the smacof package in R. The output consisted of two matrices, a 4 stimulus tones by *n* dimensions matrix of coordinates represented visually in a 'group stimulus space', (see Fig.7), and two matrices of weights (one for pretest and one for posttest) for each participant.

4.2 Cue-weighting for tone perception—INDSCAL analyses

Based on the scree plot (the plot of stress values as a function of dimensionality where the stress value is the estimation of badness of fit) and interpretability, all INDSCAL analyses generated the best dimension solution with two dimensions as there was a sharp stress decrease from a one dimensional solution to a two dimensional solution and the two dimensions can be interpreted as pitch direction and pitch height. In the following sections, we report the two dimensional group configurations of the native

control, non-native control groups and other four trainee groups. Also, individual differences in terms of cue-weighting are reported as well.

4.2.1 Cue-weighting difference between native Chinese speakers and native English speakers in the pretest

First, we replicated the results of previous cross-linguistic studies (e.g., Gandour 1983; Chandrasekaran et al. 2010) on cue-weighting for tone perception. Fig. 7 shows the group configuration of native Chinese speakers' cue-weighting on two dimensions. On Dimension 1 (Dim 1), T2 and T4 were judged to be the most distant whereas on Dimension 2 (Dim 2), T1 and T3 were judged to be the most distant. T2 is a rising tone and T4 is a falling tone, thus, Dim 1 can be interpreted as pitch direction. T1 is a high level tone and T3 is a low dipping tone, thus, Dim 2 can be interpreted as pitch height. In INDSCAL, Dim 1 accounts for more variance than Dim 2 and any higher dimensions. It means Dim 1 is weighted more than Dim 2 in terms of judging similarity between stimuli. To compare with native Chinese speakers' cue-weighting for tone perception, we randomly selected two native English speakers from each experiment and used their RT in the speeded AX discrimination task in the pretest to examine whether their cue-weighting is different from native Chinese speakers. The ten native English speakers' group configuration is shown in Fig. 8 in which T1 and T3 were judged to be the most distant on Dim 1 whereas T2 and T4 were judged to be the most distant on Dim 2. Thus, Dim 1 can be interpreted as pitch height and Dim 2 can be interpreted as pitch direction. Thus, the native English speakers' group configuration is the opposite to the native

Chinese speakers' group configuration. Therefore, the result in terms of native Chinese speakers and native English speakers' cue-weighting for Mandarin Chinese tone perception in the current study is consistent with the results found in the previous studies on cue-weighting in tone perception (e.g., Gandour 1983; Chandrasekaran et al 2010; Huang 2001), namely, native Chinese speakers as a group weighted pitch direction more than pitch height whereas native English speakers as a group weighted pitch height more than pitch direction.

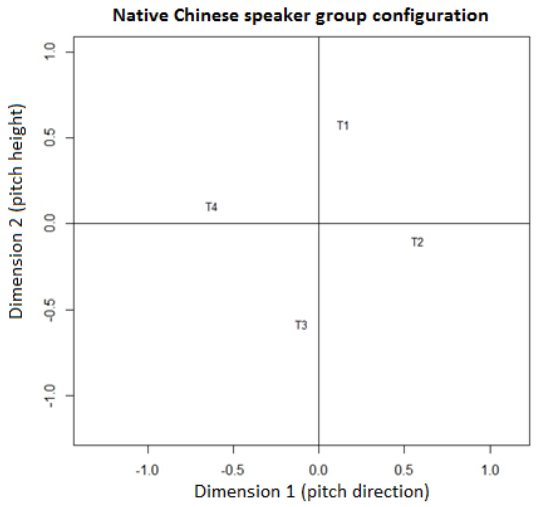

Figure 7. Group stimulus space configuration of native Chinese speakers in Omnibus INDSCAL

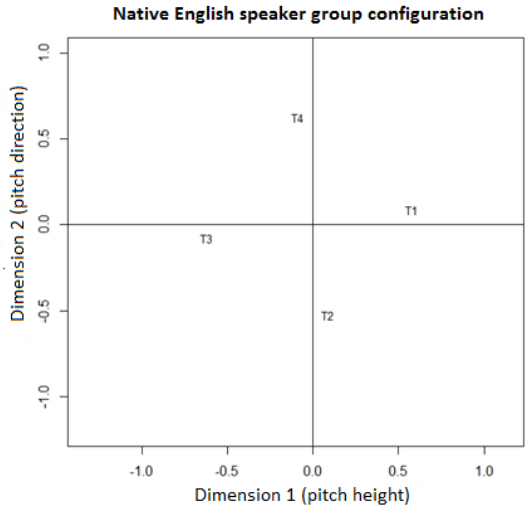

Figure 8. Group stimulus space configuration of native English speakers in Omnibus INDSCAL

The individual variability in terms of cue-weighting within the native Chinese speakers and native English speakers are illustrated in Fig. 9 and Fig. 10, respectively. In Fig. 9, the bisector indicates equal weighting on Dim 1 (pitch direction) and Dim 2 (pitch height). Anyone who is above the bisector weights Dim 2 (pitch height) more than Dim 1

(pitch direction) whereas anyone who is below the bisector weights Dim 1 (pitch direction) more than Dim 2 (pitch height). The result suggested the existence of individual variability. By visual inspection, three participants weighted Dim 1 more than Dim 2. Four participants weighted Dim 2 (pitch height) more than Dim 1 (pitch direction). Three participants weighted the two dimensions equally. However, the result did not mean that native Chinese speakers as a language group weighted pitch height more than pitch direction as more participants weighted Dim 2 (pitch height) more than Dim 1 (pitch direction). The result only indicated that there was individual variability among the native Chinese speakers in terms of cue-weighting on pitch height and pitch direction. Each individual's perceptual distance among the four tones can be calculated by multiplying the individual's cue-weighting value on each dimension and the coordinates of the group configuration. In that way, regardless the individual differences in terms of cue-weighting, most of the native Chinese speakers still perceived T2 and T4 to be the most distant on Dim 1 (pitch direction) and T1 and T3 the most distant on Dim 2 (pitch height). Similarly, most of the native English speakers still perceived T1 and T3 to be the most distant on Dim 1 (pitch height) and T2 and T4 the most distant on Dim 2 (pitch direction).

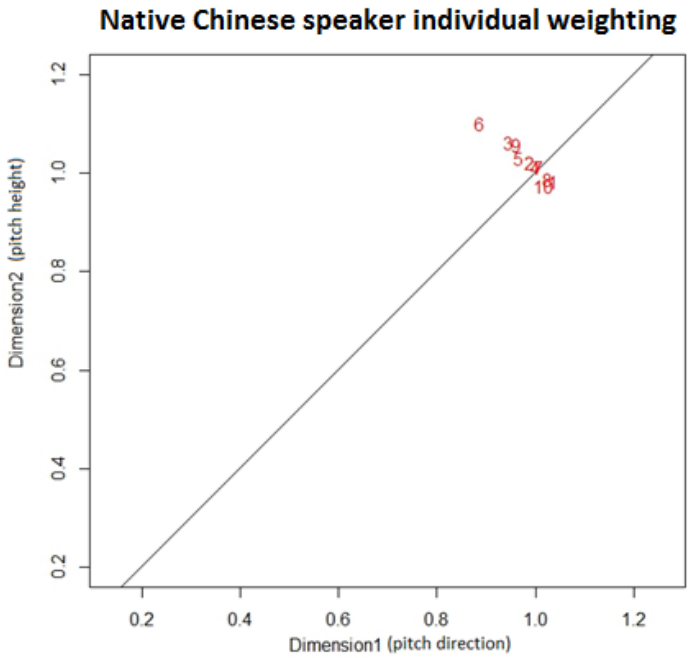

Figure 9. Normalized cue-weighting coefficients on Dim 1 and Dim 2 of ten native Chinese speakers. Dim 1 corresponds to pitch direction and Dim 2 corresponds to pitch height. The digits correspond to subject ID.

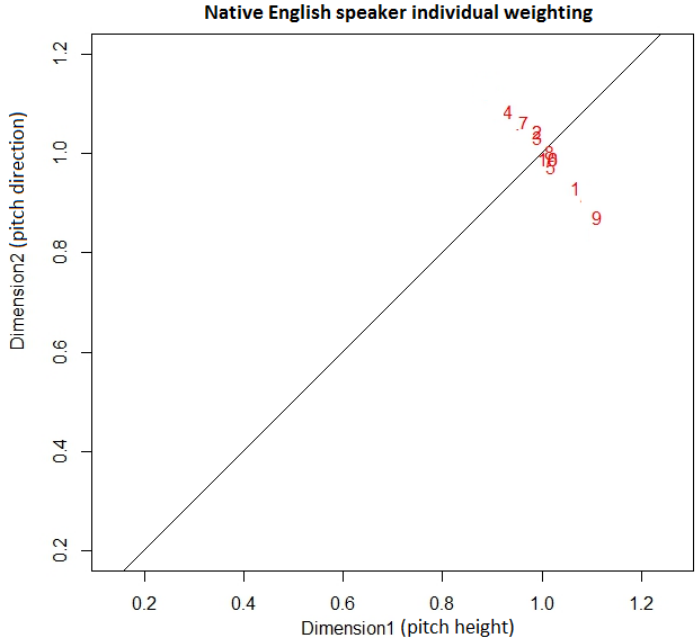

Figure 10. Normalized cue-weighting coefficients on Dim 1 and Dim 2 of ten native English speakers randomly selected from the 50 native English speakers. Dim 1 corresponds to pitch height and Dim 2 corresponds to pitch direction. The digits correspond to subject ID.

When the ten native Chinese speakers and the ten native English speakers were pooled together for the INDSCAL analysis, T1 and T3 were the most distant on Dim 1 whereas T2 and T3 were the most distant on Dim 2 in the group configuration map shown in Fig. 11. Thus, Dim 1 can be interpreted as pitch height whereas Dim 2 can be interpreted as pitch direction. The difference between the native Chinese speaker group and the native English speaker group in terms of perception of the four different lexical tones is reflected in the individual cue-weighting result. The individual cue-weighting result showed that the majority of the native Chinese speakers weighted pitch direction more than pitch height while the majority of the native English speakers weighted pitch height more than pitch direction, as Fig. 12 shows.

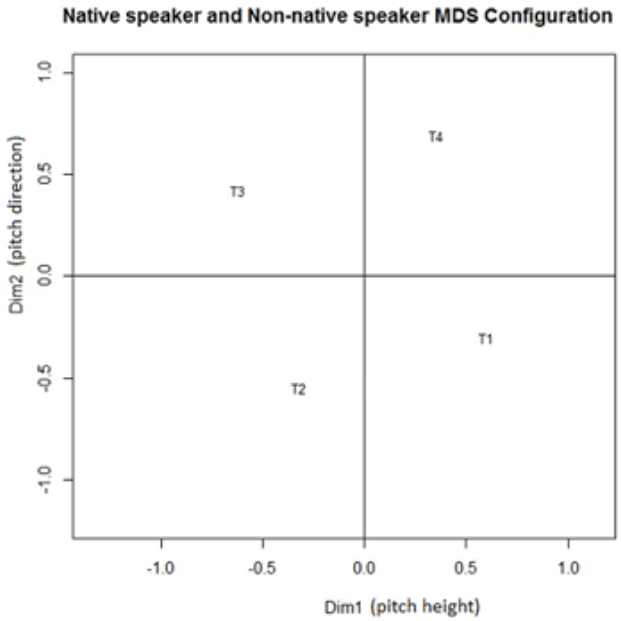

Figure 11. Group stimulus space configuration of omnibus INDSCAL analysis when ten native Chinese speakers and ten native English speakers were pooled together.

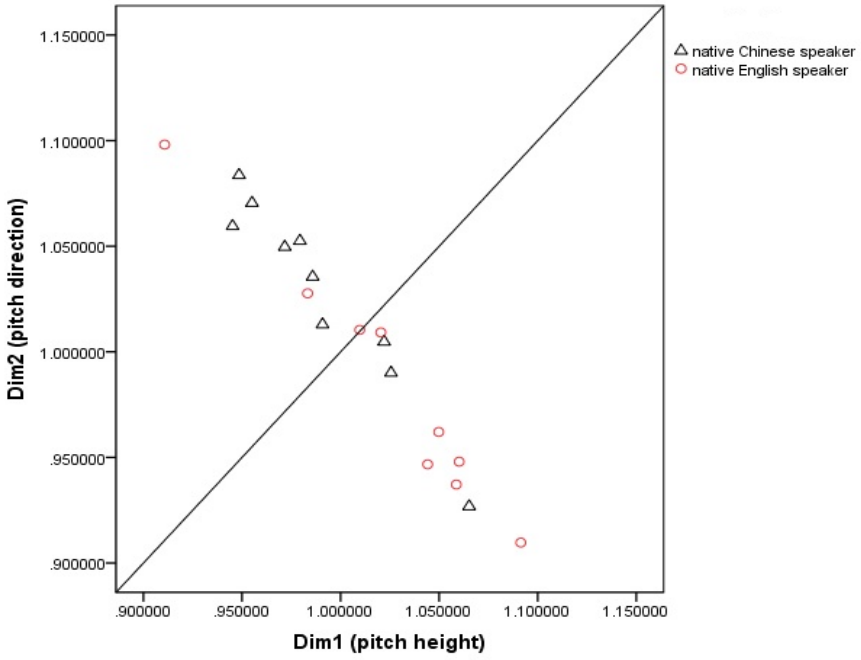

Figure 12 Individual cue-weighting on pitch height and pitch direction for native Chinese speakers and native English speakers.

As Fig. 12 shows, seven out of ten native Chinese speakers weighted Dim 2 (pitch direction) more than Dim 1 (pitch height) and seven out of ten native English speakers weighted Dim 1 (pitch height) more than Dim 2 (pitch direction). Based on the group configuration and individual cue-weightings, we can argue that native Chinese speakers as a language group weighted pitch direction more than pitch height whereas native English speakers as a language group weighted pitch height more than pitch direction. However, individual variability existed among both language groups as three native Chinese speakers weighted pitch height more than pitch direction and three native English speakers weighted pitch direction more than pitch height.

The relation between the group configuration and the individual cue-weighting preferences can be interpreted as follows: the individual cue-weightings on Dim 1 (pitch height) and Dim 2 (pitch direction) stretch the space so that the distances among the four lexical tones estimated by the INDSCAL algorithm based on 1/RT reflect their psychometric distances in the perceptual space of the participants as a group. The larger distance between T2 and T4 on Dim 1 in the group configuration map of the native Chinese speakers reflected that the native Chinese speakers as a group judged T2 and T4 overall as the most different tone pair. Though T1 and T3 were judged with a larger distance on Dim 2, the overall perceptual difference between T1 and T3 was still not as large as that between T2 and T4. Thus, on one hand, the group configuration in INDSCAL reflects which tone pair is typically judged to be the most distant by a group of participants (e.g., native Chinese speakers), on the other hand, the model captures the individual differences. Another way to look at the individual differences is equivalent to specifying how idiosyncratically a participant behaves from the most typical behavior in the group. The bisector also indicates the most typical cue-weighting scenario within a language group. The further away from this bisector, the more idiosyncratically a participant is in terms of the cue-weighting on the two dimensions.

4.2.2 Cue-weighting results of monosyllable training groups

In this section, we report the cue-weighting results of the non-native control group and two trainee groups, one of which was trained with variance manipulated

monosyllables (Exp. 1b) whereas the other was trained with multi-talker monosyllables (Exp. 1c).

The pretest group configurations of the non-native control, variance manipulated training and multi-talker training groups are illustrated in Fig. 13a, Fig. 13b, and Fig. 13c, respectively. Their corresponding posttest group configurations are shown in Fig. 14a, Fig. 14b and Fig. 14c respectively. In the pretest, all three groups consistently showed that T1 and T3 were the most distant on Dim 1 whereas T2 and T4 were the most distant on Dim 2. This was just the opposite pattern compared to the group configuration of the native Chinese speakers. It indicated that English speakers as a group in the pretest primarily depended on pitch height to judge the similarity among different lexical tones. This result was consistent with previous studies on English speakers' perception of lexical tones (Gandour 1983, Huang 2001, Chandrasekaran et al. 2010).

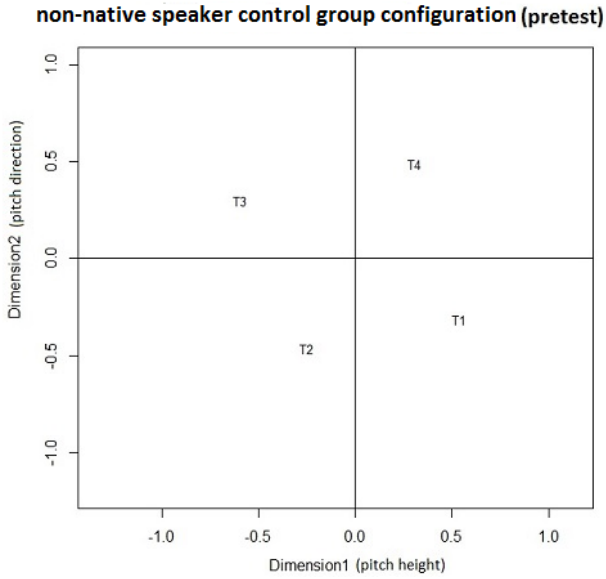

Figure 13a

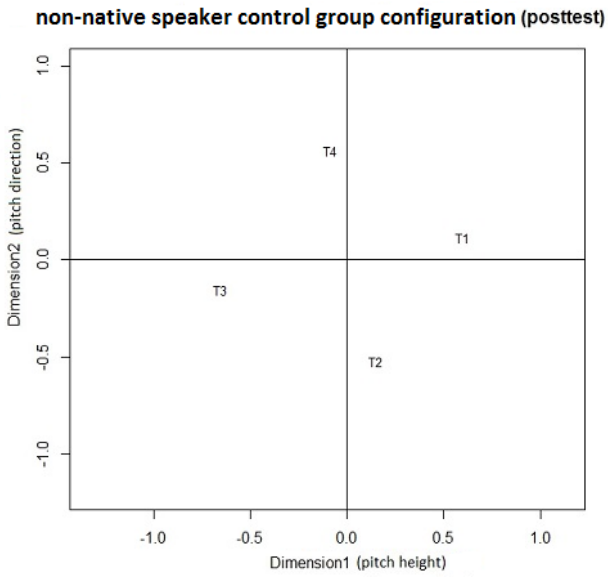

Figure 14a

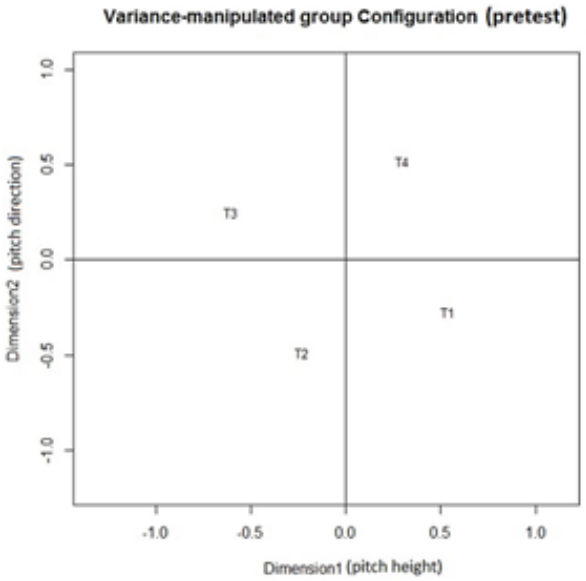

Figure 13b

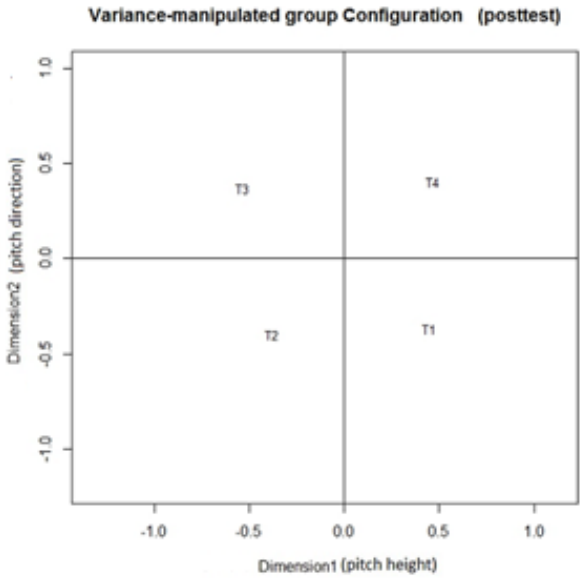

Figure 14b

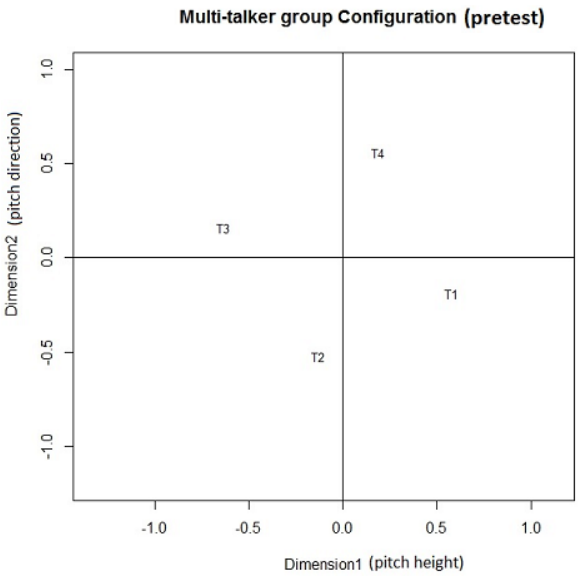

Figure 13c

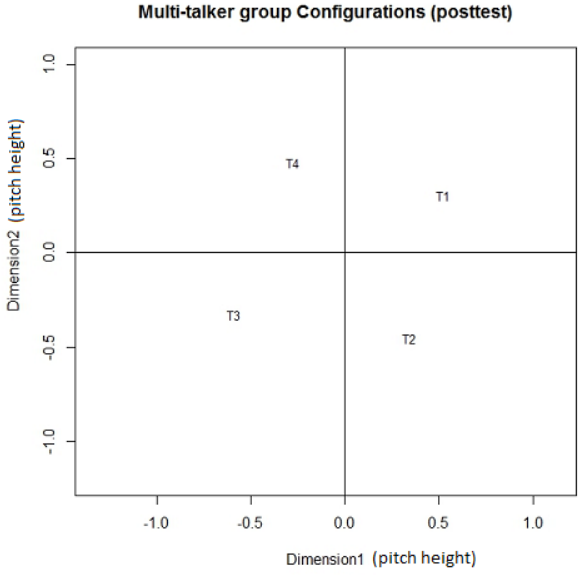

Figure 14c

Figure 13&14. Figure 13(a)-13(c) illustrate the non-native control, the variance manipulated and the multi-talker training conditions' group configuration maps of four Mandarin Chinese tones in the pretest. Figure 14(a)-14(c) illustrate the three groups' configuration maps in the posttest.

In the posttest, after four days of video-game training, the group configuration of the non-native control group as shown in Fig. 14a did not change much as Dim 1 was still pitch height as it was the dimension on which T1 and T3 were judged to be the most different, and Dim 2 was still pitch direction in that it was the dimension on which T2 and T4 were judged to be the most different. However, the group configuration of the variance manipulated training group and multi-talker training group seemed to have undergone some change. As Fig. 13b and 14b show, the perceived distances among the four lexical tones for the variance manipulated training group seemed to have turned clockwise, whereas as Fig. 13c and 14c show, the perceived distances among the four lexical tones for the multi-talker training group seemed to have turned counter-clockwise. Assuming Dim 1 and Dim 2 did not change qualitatively within such a short period of training, either direction of turning suggested that a reassignment of the weights on the two dimensions had occurred. More specifically, after the training, the variance manipulated training group perceived T2 and T4 to be more different on Dim 1 and T1 and T3 to be more different on Dim 2, as shown in Fig. 14b. A similar result was found for the multi-talker training group as shown in Fig. 14c. The pretest and posttest group configuration change in the variance manipulated and multi-talker training groups suggested that cue-weighting had been shifted towards pitch direction. It is worth noting that the perceived distance among the four tones of the non-native control group also turned counter-clockwise after playing the video game without any lexical tone input, as shown in Fig. 13a and 14a. However, a close look showed that in the posttest configuration as shown in Fig.14a, the distance between T2 and T4 became smaller on

Dim 1 (pitch height) but the distance between T1 and T3 was still large on Dim 1 (pitch height), indicating more weights had been assigned to Dim 1 (pitch height). This pattern was opposite to the cue-weighting shift that occurred to the variance-manipulated and multi-talker training groups. The rotation of the group configurations of the variance manipulated training and multi-talker training groups suggested that these two trainee groups' cue-weighting became more nativelike.

The pretest individual cue-weightings of the non-native control group, variance manipulated group and multi-talker group are illustrated in Fig. 15a, Fig. 15b, and Fig. 15c respectively. Their corresponding posttest individual cue-weightings are shown in Fig. 16a, Fig. 16b and Fig. 16c respectively. By comparing the individual cue-weightings before and after the training, we can see that both the variance manipulated and multi-talker training conditions made all the individuals shift more cue-weighting towards Dim 2 (pitch direction) whereas the non-native control group did not show any trend of cue-weighting shift towards Dim 2 (pitch direction). Before the training, it seemed that the participants in the multi-talker training group had a larger variability than the variance-manipulated training group in terms of individual cue-weighting. After the training, the variability of individual cue-weighting within the multi-talker training group was reduced as almost all the participants weighted Dim 2 (pitch direction) more than Dim 1 (pitch height).

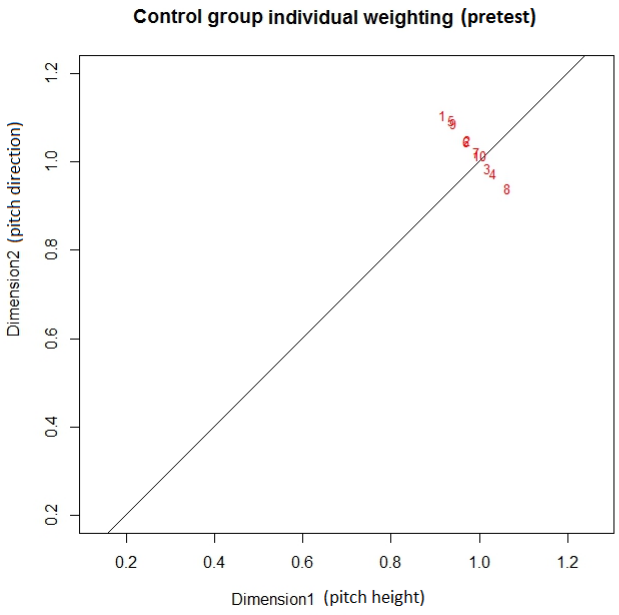

Figure 15a

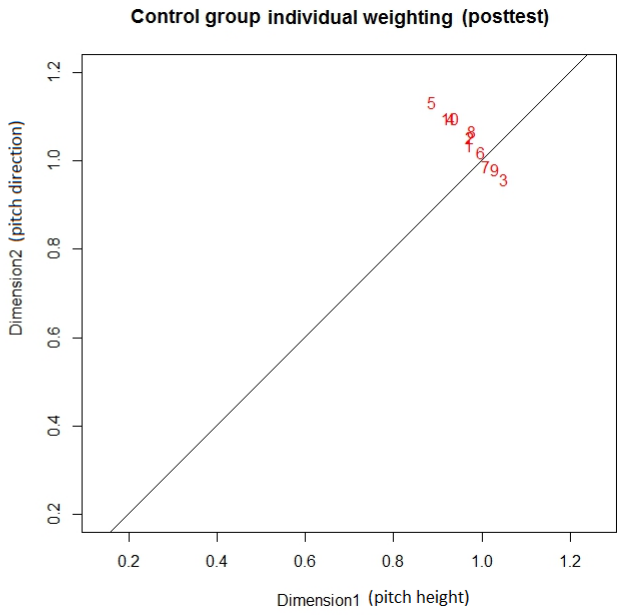

Figure 16a

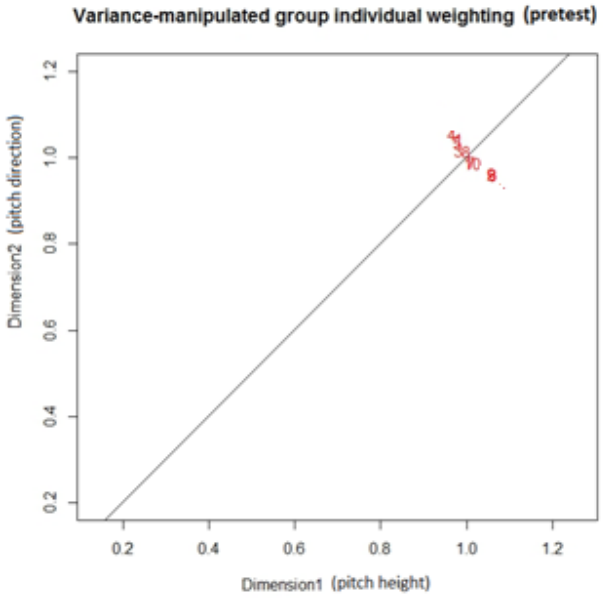

Figure 15b

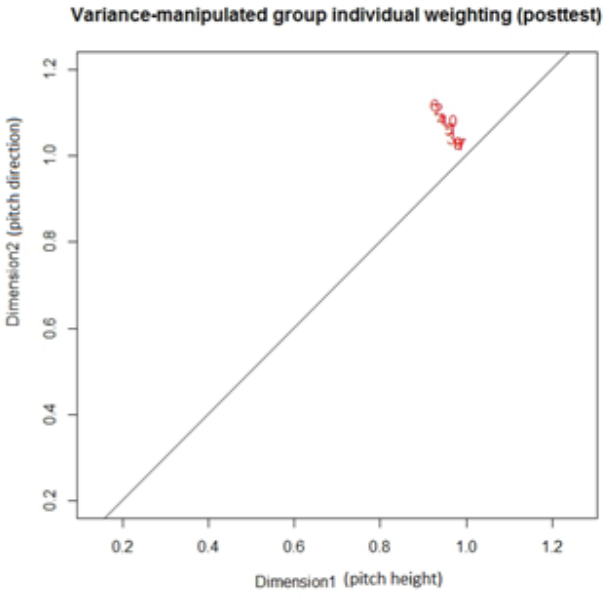

Figure 16b

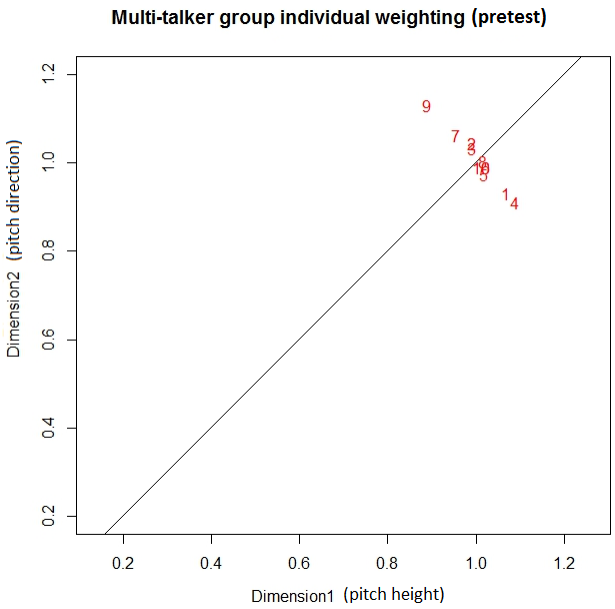

Figure 15c

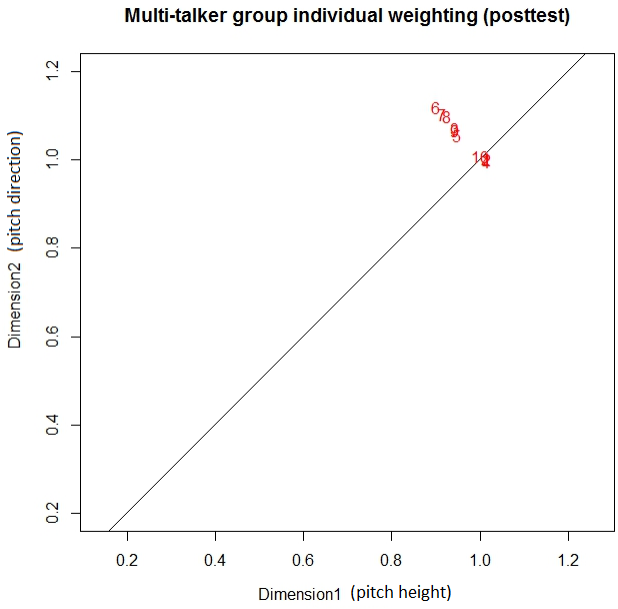

Figure 16c

Figure 15 & 16. Figure 15(a)-15(c) illustrate the individual weightings of non-native control, variance-manipulated and multi-talker training groups in the pretests. Figure 16(a)-16(c) illustrate the individual weightings of the three conditions in the posttests.

To examine the individual cue-weightings quantitatively, we conducted a 2x3 repeated measures ANOVA (within-subject: Test (pretest vs. posttest); between-subject: Experiments (Exp.1a—monosyllables without tone training/non-native control, Exp.1b—variance-manipulated monosyllable training, and Exp.1c—multi-talker monosyllable training) using cue-weighting values on Dim 1 (pitch height) and Dim 2 (pitch direction) as the Dependent Variable (henceforth DV).

In terms of cue-weights on Dim 1 (pitch height), the result showed a main effect of Test ($F(1,27)=6.28$, $p<.05$) and a significant Text x Training interaction ($F(2,27)=1.77$,

p<.05). Fig. 17 shows the non-native control group and two monosyllable training groups' cue-weights on Dim 1 (pitch height) before and after the training.

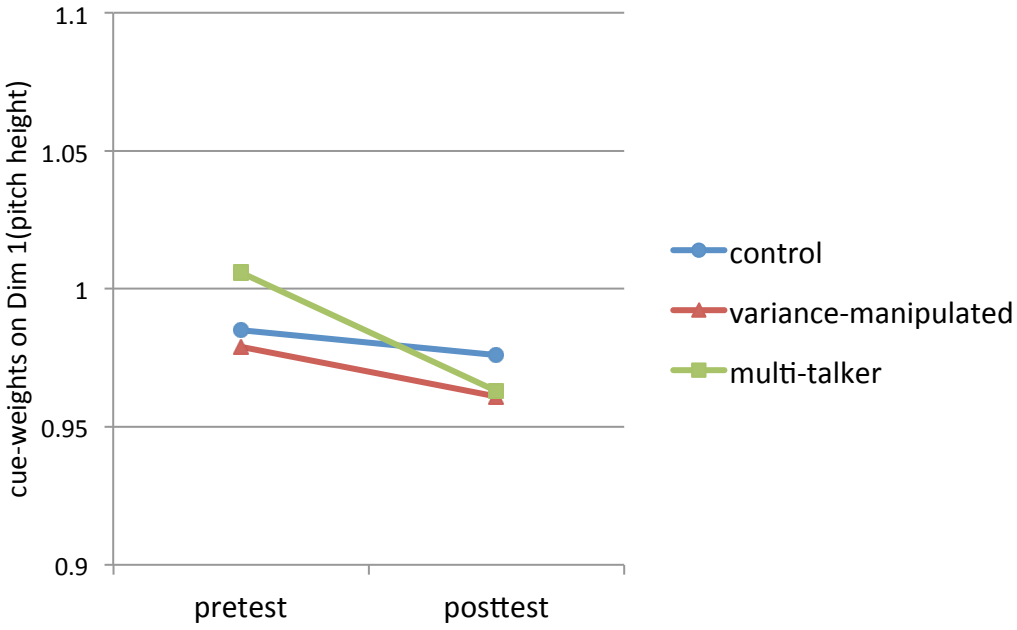

Figure 17. The non-native control group and two monosyllable training groups' cue-weights on Dim 1 (pitch height) in pretest and posttest.

As Fig. 17 shows, the multi-talker training group had the largest cue-weight decrease on the pitch height dimension relative to the other two groups. To examine the cue-weights on Dim 1 (pitch height) within each training group, we examined the simple effect of Test. The results showed that only the multi-talker training group had a significant cue-weighting decrease on the pitch height dimension (F(1,27)=7.1, p<.05) whereas the variance manipulated training group and the non-native control group did not have a cue-weighting decrease on the pitch height dimension. Fig. 18 illustrates the simple effect of Test toward the cue-weighting on pitch height within each group.

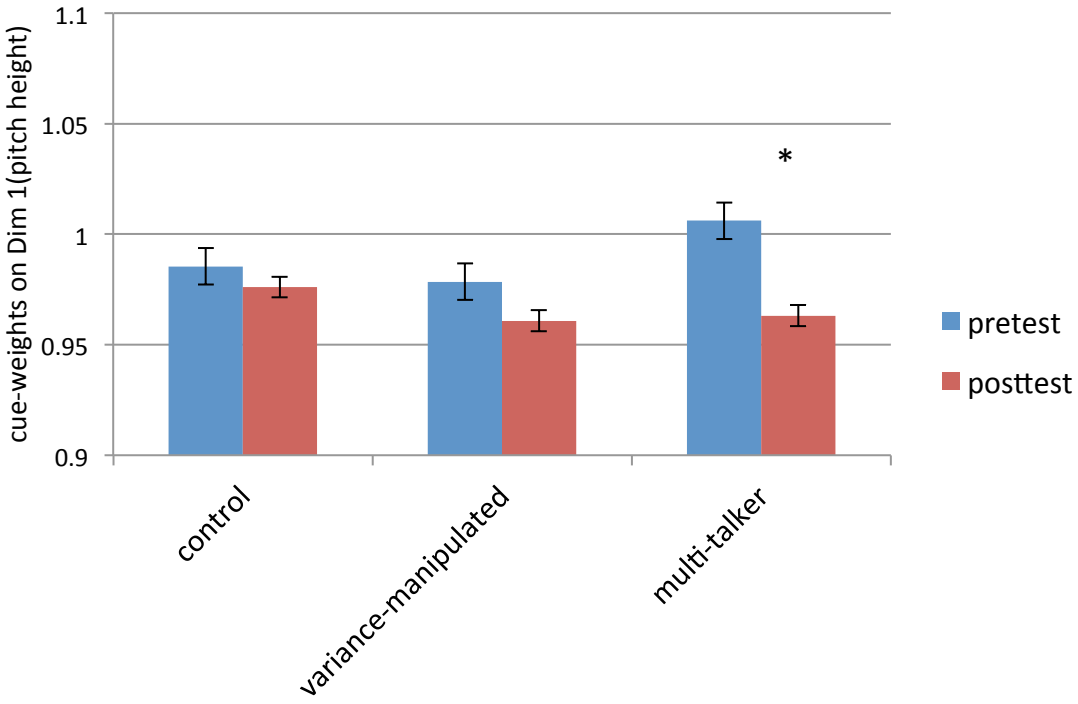

Figure 18. The simple effect of Test on the cue-weighting on Dim 1 (pitch height) in the control group and the two monosyllable training groups. * indicates $p < .05$.

In terms of cue-weights on Dim 2 (pitch direction), the result showed a main effect of Test ($F(1,27)=5.7$, $p<.05$) and a significant Text x Training interaction ($F(2,27)=2.82$, $p<.05$). Fig. 19 shows the control group and two monosyllable training groups' cue-weights on Dim 2 (pitch direction) before and after the training.

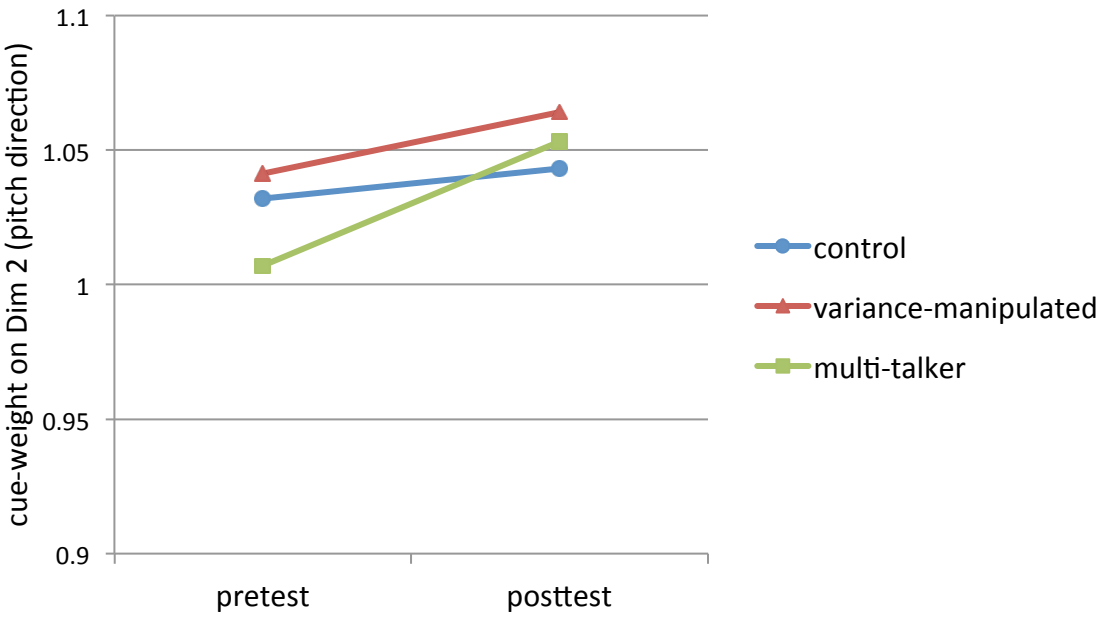

Figure 19. The non-native control group and two monosyllable training groups' cue-weights on Dim 2 (pitch direction) in pretest and posttest.

As Fig. 19 shows, the multi-talker training group had the largest cue-weight increase on the pitch direction dimension relative to the other two groups. To examine the cue-weights on Dim 2 (pitch direction) within each training group, we examined the simple effect of Test. The results showed that both the variance-manipulated training and multi-talker training groups had a significant cue-weighting increase on the pitch direction dimension (variance-manipulated training group: $F(1,27)=1.44$, $p<.05$; multi-talker training group: $F(1,27)=5.28$, $p<.05$) whereas the non-native control group did not have a cue-weighting increase on pitch direction. Fig. 20 illustrates the effect of Test on each group's cue-weighting on pitch direction.

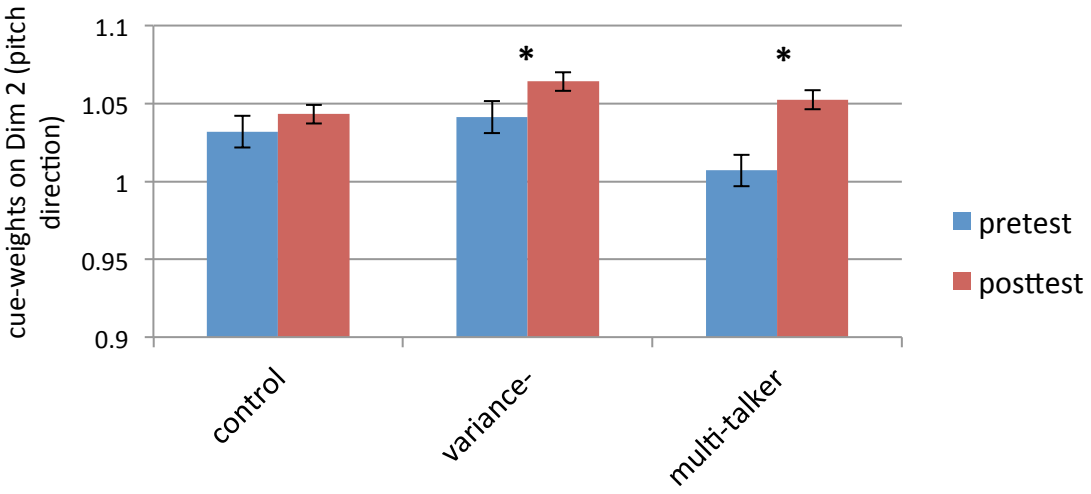

Figure 20. The simple effect of Test on the cue-weighting on Dim 2 (pitch direction) in the non-native control group and the two monosyllable training groups. * indicates $p < .05$.

One point worth mentioning here is that although the variance-manipulated training successfully helped naïve listeners shift their cue-weighting towards pitch direction after the training, we cannot make a strong claim about the effect of variance on shifting cue-weighting. The reason is that although the variance on pitch direction was smaller than that on pitch height in terms of jnd in our training stimuli, it is still possible that the theoretical just noticeable difference between two tone tokens for the same lexical tone category cannot be heard by the naïve listeners,as the jnd for discriminating the synthesized pitch contours found in the psychophysics studies may not fully apply to the discrimination of naturally produced pitch contours. Thus, we need to be cautious about claiming that the smaller variance on pitch direction made naïve listeners shift their cue-weighting towards pitch direction.

Another point worth mentioning is that the overlap on the pitch direction dimension among the training tokens in the multi-talker training seemed not to hamper the cue-weighting shift towards pitch direction. Though there was no overlap on the pitch

direction dimension among the training tokens in the variance-manipulated training, its training effect was not as robust as the multi-talker training in terms of shifting cue-weighting towards pitch direction. Thus, it seemed that the overlap on pitch direction to certain extent among the training tokens did not hamper the cue-weighting shift at all. It somehow suggested that sound categorization in a multi-dimensional acoustic space may never need to have a dimension that sound categories are completely distinct from each other. An optimal sound classification should allow an overlap between sound categories in the acoustic space. More on the current cue-weighting result's theoretical implications for sound categorization is discussed in Chapter Five.

4.2.3 Summary of cue-weighting results of monosyllable training conditions

The INDSCAL analysis generated two-dimensional configurations that best reflected the perceptual distance among the four lexical tones within the native Chinese speaker group, the non-native control group and the two monosyllable training groups. The native Chinese speaker group weighted pitch direction more than pitch height whereas the three English speaker groups weighted pitch height more than pitch direction. Regardless of the individual variability of cue-weighting on the two dimensions as shown in Figs.15 and 16, both the variance-manipulated training group and the multi-talker training group shifted cue-weights towards pitch direction after the training whereas the non-native control group did not show any cue-weighting shift towards pitch direction. Moreover, the multi-talker training had a greater cue-weighting increase on pitch direction relative to the variance-manipulated training group. Interestingly, only the multi-talker training group showed a cue-weighting decrease on the pitch height

dimension whereas the other two groups did not show any cue-weighting decrease on the pitch height dimension. These results suggested that multi-talker training may be more robust than variance-manipulated training in terms of boosting the cue-weighting on the most reliable acoustic cue and at the same time reducing the cue-weighting on the secondary or less reliable acoustic cue. Chapter Five discusses more implications regarding the cue-weighting results.

4.2.4 Cue-weighting results of disyllable training conditions

Two groups of native English speakers were trained on disyllables where the second syllables were the variance-manipulated tone tokens used in Experiment 1b. The first disyllable training group was trained with disyllables that were minimal pairs (Exp. 2a) whereas the second disyllable training group was trained with disyllables that were non-minimal pairs (Exp. 2b).

The pretest group configurations of the minimal pair disyllable training group and non-minimal pair disyllable training group are illustrated in Fig. 21a, and Fig. 21b respectively. Their corresponding posttest group configurations are shown in Fig. 22a, Fig. 22b, respectively. Same as the monosyllable training groups, the two disyllable training groups both showed that T1 and T3 were the most distant on Dim 1 whereas T2 and T4 were the most distant on Dim 2. Thus, Dim 1 in the group configurations of the two disyllable training groups can be interpreted as pitch height and Dim 2 can be interpreted as pitch direction. In the posttest, the group configurations of the two disyllable training groups were still quite similar to the ones in the pretest.

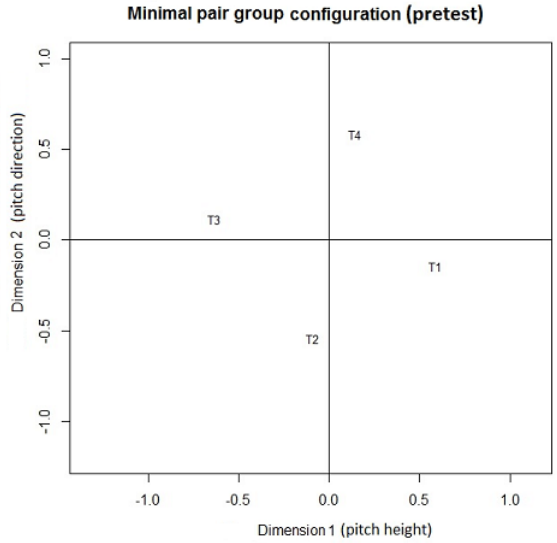

Figure 21a

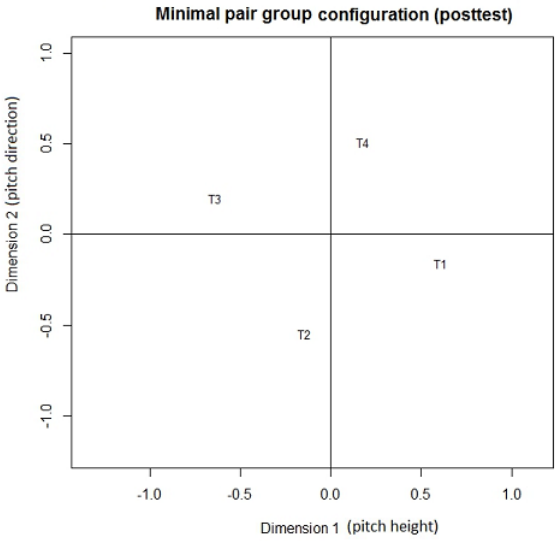

Figure 22a

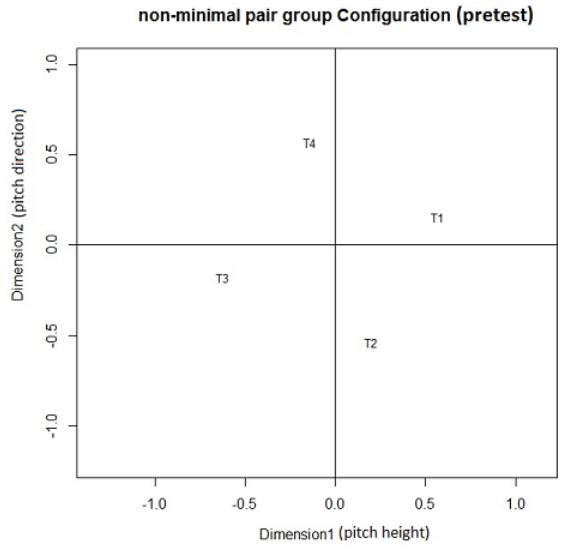

Figure 21b

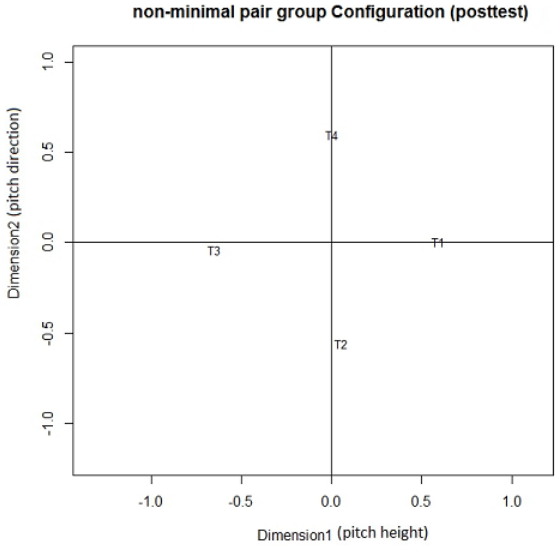

Figure 22b

Figure 21&22. Figure 21(a) and 21(b) illustrate the group configuration in the pretests of the disyllable minimal pair and disyllable non-minimal pair training conditions. Figure 22(a)-22(b) illustrate the two training conditions' group configuration in the posttests.

The individual cue-weightings of the two disyllable training groups in the pretest are shown in Fig. 23a and 23b. Their corresponding posttest individual cue-weightings are shown in Fig. 24a and 24b. There was no clear trend of cue-weighting shift towards either dimension after the training for both disyllable training groups.

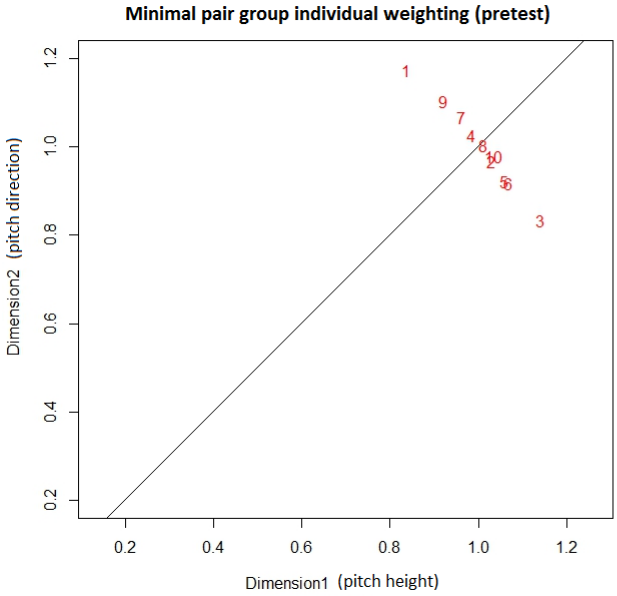

Figure 23a

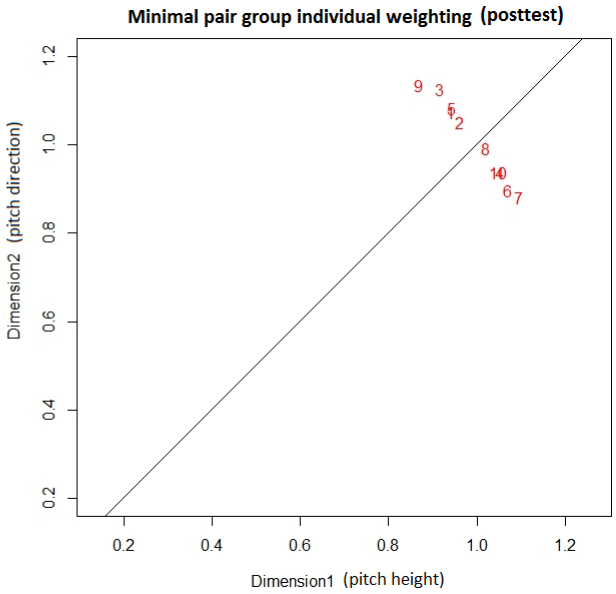

Figure 24a

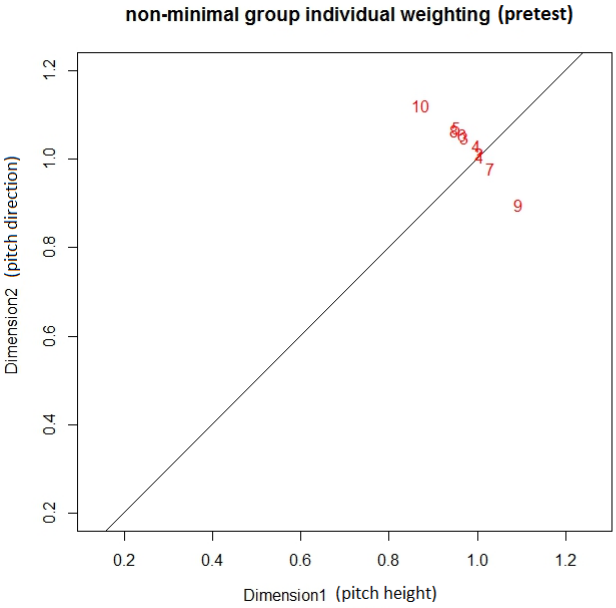

Figure 23b

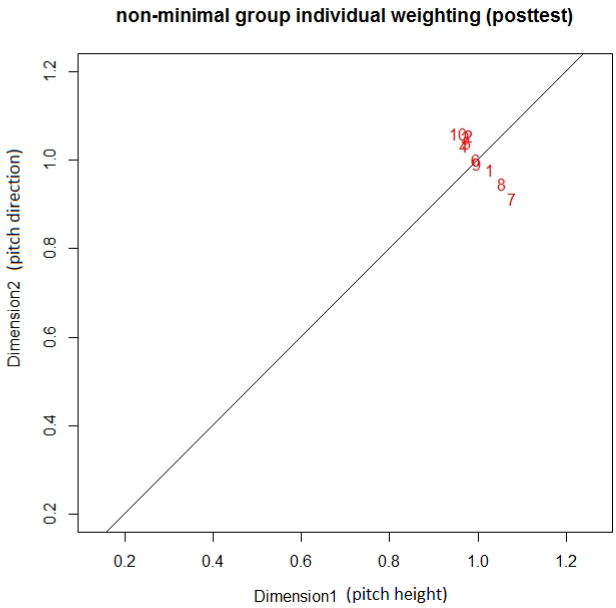

Figure 24b

Figure 23 & 24. Figure 23(a) and 23(b) illustrate the individual weighting of the disyllable minimal pair and disyllable non-minimal pair training conditions in the pretests. Figure 24(a) and 24(b) illustrate the individual weighting of the two training conditions in the posttests.

Same as the monosyllable training conditions, we conducted a 2x3 repeated measures ANOVA (within-subject: Test (pretest vs. posttest); between-subject: Experiments (Exp. 1b—variance-manipulated monosyllable training/non-native control, Exp. 2a—variance-manipulated minimal disyllable pair training, and Exp. 2b—variance-manipulated non-minimal disyllable pair training) using the cue-weighting values on Dim 1 (pitch height) and Dim 2 (pitch direction) as DVs. We aimed to examine whether the benefits of variance-manipulated training in terms of shifting cue-weight toward pitch direction can also be found by using disyllabic training stimuli.

In terms of the cue-weighting on Dim 1 (pitch height), there was neither a significant main effect nor a significant interaction. The result suggested that none of the three groups had cue-weighting change on the pitch height dimension.

In terms of the cue-weighting on Dim 2 (pitch direction), there was no significant main effect of Test or Training; however, there was a significant Test x Training interaction ($F(2,27)=2.69$, $p<.05$), as shown in Fig. 25.

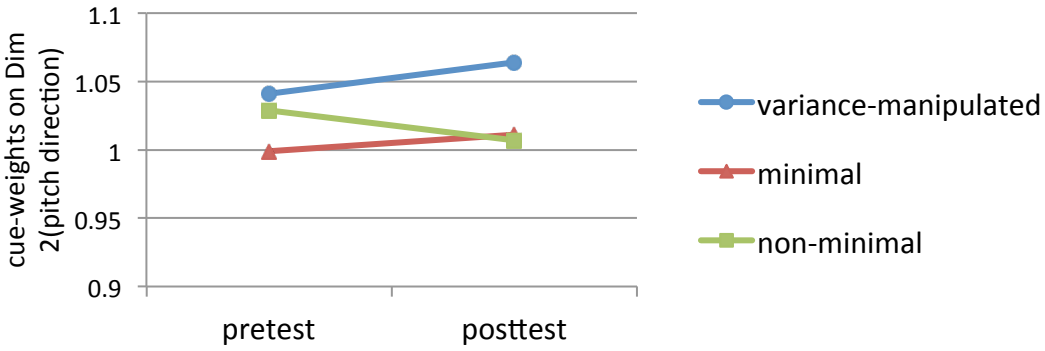

Figure 25. The variance-manipulated training group and two disyllable training groups' cue-weights on Dim 2 (pitch direction) in pretest and posttest.

Since there was a significant Test x Training interaction, we examined the simple effect of Test on each training condition to see if there was a significant cue-weighting shift towards the pitch direction dimension in any of the three training groups. The results showed that only the variance manipulated training group had a significant cue-weighting increase on the pitch direction dimension ($F(1,27)=2.66$, $p<.05$) whereas the two disyllable training groups did not have any cue-weighting change on the pitch direction dimension. Fig. 26 shows the simple effect results.

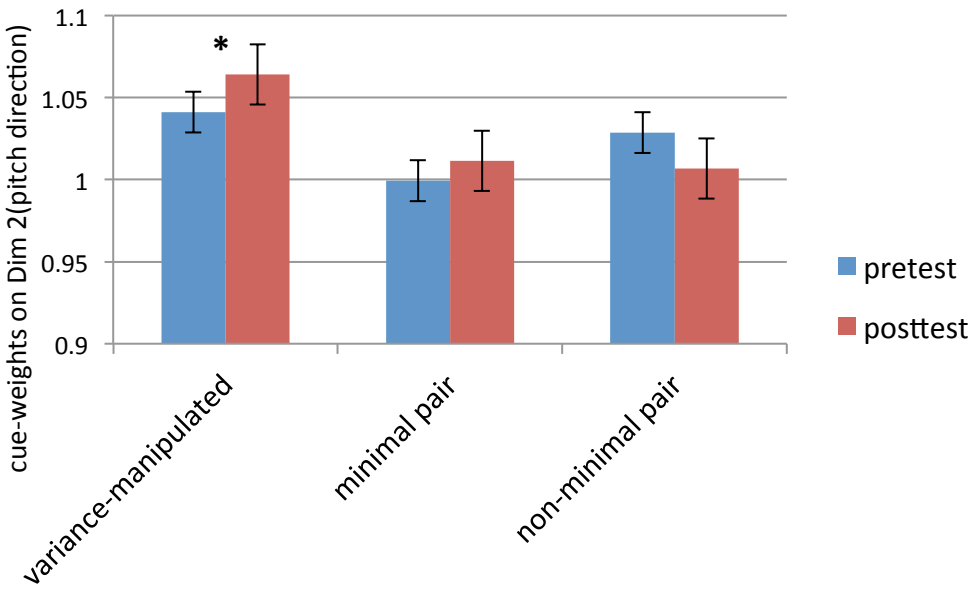

Figure 26. The simple effect of Test on the cue-weighting on Dim 2 (pitch direction) in the variance manipulated training group and the two disyllable training groups. * indicates $p < .05$.

4.2.5 Summary of cue-weighting results of disyllable training conditions

The INDSCAL analyses generated two dimensional configurations as the best solution that reflected the perceptual distance among four lexical tones within the minimal pair disyllable and non-minimal pair disyllable training groups. Dim 1 can be interpreted as pitch height whereas Dim 2 can be interpreted as pitch direction. Both disyllable training groups weighted pitch height more than pitch direction. After four days of training, neither group showed a cue-weighting shift towards pitch direction. These results suggested that adding a preceding syllable to the variance-manipulated syllables did not help the participants shift cue-weighting towards pitch direction at all.

More importantly, adding such a preceding syllable blocked the effect of the variance-manipulated monosyllable in shifting cue-weights towards pitch direction.

4.3 Sensitivity to lexical tones in monosyllables and disyllables

The first speeded AX discrimination task (i.e. tone discrimination in monosyllables) and the second non-speeded AX discrimination task (i.e. tone discrimination in disyllables) are used to examine how sensitive the naïve listeners were to the tone differences in different contexts. To measure their sensitivity, we calculated the d' score for each participant before and after the training. Native Chinese speakers achieved ceiling results for the tone discrimination both in the monosyllables (mean accuracy rate: 99%, mean d'=4.5) and in the disyllables (mean accuracy rate: 97.8%, mean d'=4.4). For naïve listeners, their tone discrimination performance in monosyllables was high in the pretest, namely, all groups of native English speakers including the non-native control group had a mean accuracy rate over 80%. However, the naïve listeners' tone discrimination performance in disyllables was much worse than that in monosyllables. We conducted a 2 x 5 repeated measures ANOVA (Within-subject: monosyllable vs. disyllable; Between-subject: 5 different training conditions), using pretest d' score as DV to examine if the d' score in disyllables was significantly lower than that in monosyllables. The result showed a significant main effect of Syllable type ($F(1,45)=759$, $p<.001$), namely, the overall d' score in disyllables was lower than that in monosyllable (mean d' in monosyllable: 3.78; mean d' in disyllable: 1.51). There was no significant Syllable x Training interaction. Thus, the result suggested that naïve listeners

consistently had lower d' score in disyllable context than that in monosyllable context. Although the naïve listeners had significantly lower d' for tone discrimination in disyllables, the accuracy rate was above chance level for the tone discrimination in disyllables in the pretest: all the trainee groups had a mean accuracy rate over 70%. The following two sections report the participants' sensitivity to lexical tones in monosyllables and disyllables respectively.

4.3.1 Sensitivity to lexical tones in monosyllables

We first conducted a 2x3 repeated measures ANOVA (within-subject: Test (pretest vs. posttest); between-subject: Experiment (Exp. 1a—monosyllables without tone training/non-native control, Exp. 1b—variance-manipulated monosyllable training and Exp.1c—multi-talker monosyllable training)) using d' as DV to examine whether participants' sensitivity to lexical tones in monosyllables improved after the monosyllable training. The results showed a significant main effect of Test ($F(1,27)=20.65$, $p<.001$) but there was not a significant Test x Training interaction. The result is shown in Fig. 27.

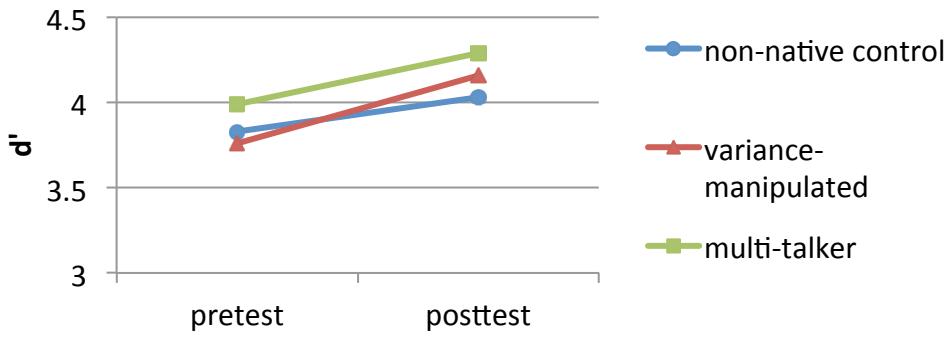

Figure 27. d' scores of the non-native control group and the two monosyllable training groups in monosyllable discrimination before and after the training.

The absence of Test x Training interaction suggested that all three groups including the non-native control group consistently had a d' increase from pretest to posttest. The result suggested that there was a practice effect for tone discrimination in monosyllables. The practice effect may come from the feedback after each trial in the pre- and posttest.

We then conducted another 2x4 repeated measures ANOVA (within-subject: Test (pretest vs. posttest); between-subject: Experiments (Exp. 1a—non-native control, Exp. 1b—variance-manipulated monosyllable training, Exp. 2a— minimal disyllable pair training that used variance-manipulated monosyllables and Exp. 2b— non-minimal disyllable pair training that used variance-manipulated monosyllables) using d' as DV to examine whether disyllable training can improve naïve listeners' sensitivity to tones in the monosyllable context. The results showed a significant main effect of Test ($F(1,36)=5.2$, $p<.05$) and a marginal significant Test x Training interaction ($F(3,36)=2.4$, $p=.06$). The result is shown in Fig. 28.

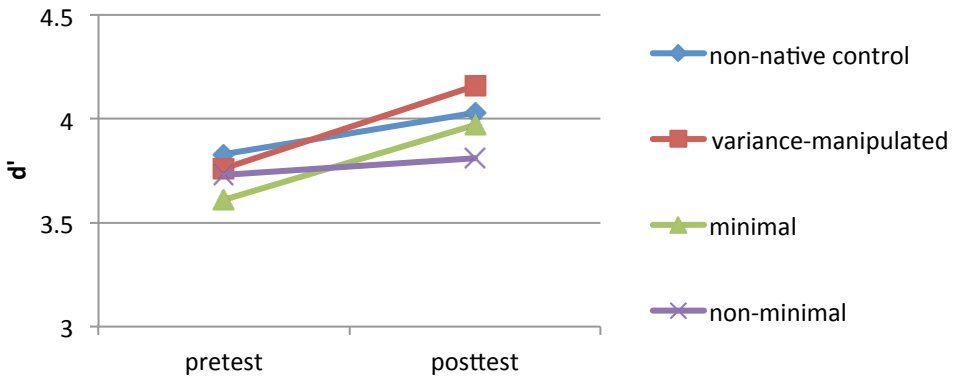

Figure 28. d' scores of the non-native control group, the variance manipulated training group, the disyllable minimal pair and disyllable non-minimal pair training groups before and after the training.

The absence of Test x Training interaction suggested that the non-native control group, the variance-manipulated training group and the two disyllable training groups consistently had a d' score increase for the tone discrimination in monosyllables after the training. However, the marginally significant Test x Training interaction suggested that there was a trend toward the non-minimal-pair disyllable training group having less d' increase than the other three groups.

4.3.2 Tone discrimination for specific tone pairs in the monosyllable context

In addition to examining the overall d' for the tone discrimination, we also calculated the d' for each tone pair in the monosyllable context. Among the 144 trials in the speeded tone discrimination task, the number of same pairs was 3 times the number of different pairs. The different tone pairs, either Tx-Ty or Ty-Tx, were repeated 12 times altogether (6 times Tx-Ty and 6 times Ty-Tx) whereas the same tone pairs Tx-Tx and Ty-Ty were repeated 36 times altogether in the experiment (18 times Tx-Tx and 18 times Ty-Ty). In order to calculate the d' for a specific tone pair, the numbers of same and different tone pairs need to be the same. To do so, we used only 12 of the 36 same tone pairs. The 12 items were randomly selected by using the RAND function in excel. To make sure the selected 12 items are not different from the remaining 24 items, we conducted a repeated-measures ANOVA using the subjects' mean accuracy rates as DV. The result showed that the subjects' accuracy rates did not differ between the two groups of items. After the selection of the 12 items, we used the same 12 items for all participants to calculate individual d' for specific tone pairs in the monosyllable context.

We conducted a 2x6x3 repeated measures ANOVA to examine the non-native control's and the two monosyllable training groups' sensitivity for discriminating specific tone pairs (Within-subject: Test (pretest vs. posttest); Tone pair (T1 vs. T2, T1 vs. T3, T1 vs. T4, T2 vs. T3, T2 vs. T4 and T3 vs. T4); Between-subject: Experiment (Exp. 1a—monosyllables without tone training/ non-native control, Exp. 1b—variance-manipulated monosyllable training and Exp.1c—multi-talker monosyllable training)), using d' as DV. The results showed a significant main effect of Test ($F(1, 125)=10.5$, $p<.01$), a significant main effect of Tone pair ($F(5, 125)=5.8$, $p<.01$) and a significant Test x Experiment x Tone pair interaction ($F(10,125)=3.9$, $p<.05$). The main effects of Test and Tone pair are illustrated in Fig. 29.

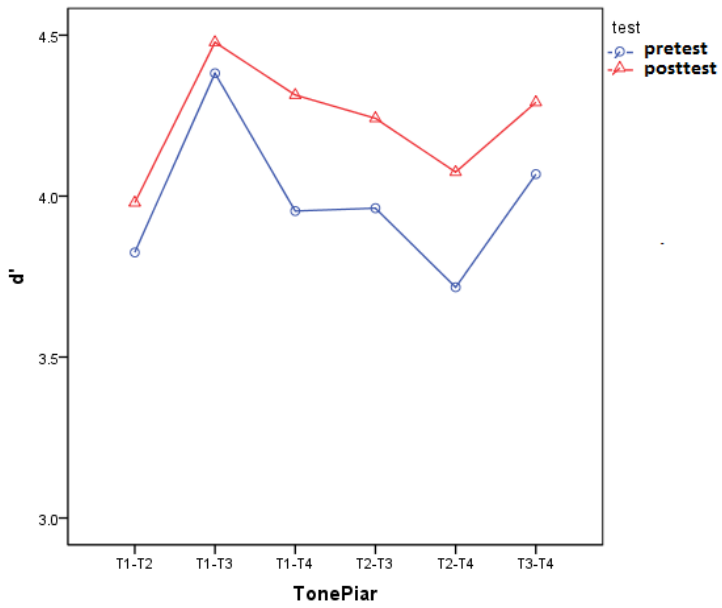

Figure 29. The d' averaged across the non-native control, variance manipulated and multi-talker training conditions in the pre- and posttest as a function of Tone pair.

Fig. 29 shows that overall, the d' in the posttest was higher than the one in the pretest. The participants had the highest d' for T1-T3 and lowest d' for T2-T4 in the pretest. The result for the pretest seemed to be related to the cue-weighting result that the native English speakers weighted pitch height more than pitch direction. Because of the higher cue-weighting on pitch height, participants were most sensitive to T1-T3 that differ the most in terms of pitch height (T1 is a high level tone; T3 is a low dipping tone) and least sensitive to T2-T4 that differ most in terms of pitch direction (T2 is a rising tone; T4 is a falling tone). After the training, the participants had the lowest d' for T1-T2 but still had the highest d' for T1-T3. The Test x Experiment x Tone pair interaction is illustrated in Fig. 30 (a)-(c).

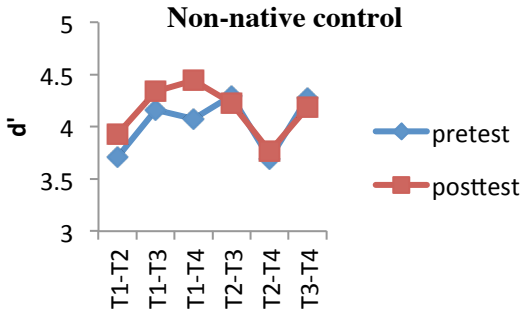

Figure 30a.

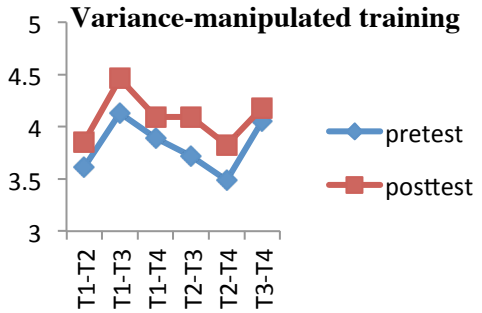

Figure 30b.

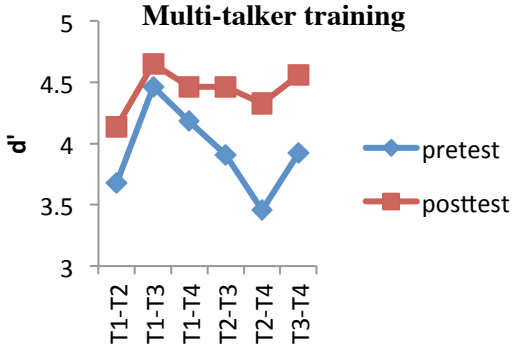

Figure 30c.

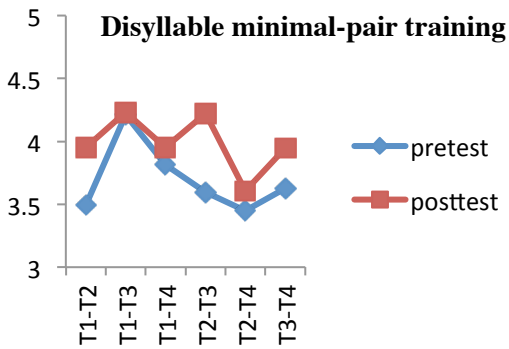

Figure 30d.

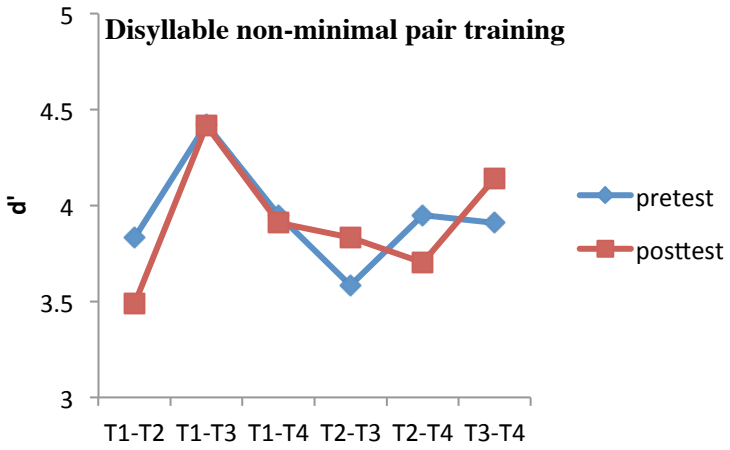

Figure 30e.

Figure 30. The d' of the non-native control, variance manipulated training and multi-talker training groups in the pre- and posttest as a function of six tone pairs.

Fig. 30(a)-(c) show that, although the non-native control, variance-manipulated and multi-talker training groups all had d' increase in the posttest compared to the pretest, the increase of d' varied across the specific tone pairs. The crucial result was that the non-native control group did not have a d' increase for T2-T4 whereas the variance-manipulated training and the multi-talker training groups had a d' increase for T2-T4 after the training. Again, the result is related to the cue-weighting shift that occurred to the variance-manipulated and multi-talker training groups, namely, both groups shifted cue-weighting towards pitch direction after the training whereas the non-native control group did not have any cue-weighting shift towards pitch direction after the training. Moreover, the multi-talker training group, which had more cue-weighting shift towards pitch direction, also had a larger d' increase for T2-T4 relative to the variance-manipulated training group. In general, it seemed that participants' cue-weighting result is related to participants' sensitivity to the difference of specific tone pairs.

We conducted another 2x6x4 repeated measures ANOVA to examine the participants' sensitivity to the difference of specific tone pairs for the disyllable training groups (Within-subject: Test (pretest vs. posttest); Tone pair (T1 vs. T2, T1 vs. T3, T1 vs. T4, T2 vs. T3, T2 vs. T4 and T3 vs. T4); Between-subject: Experiment (Exp. 1a—monosyllables without tone training/non-native control, Exp. 1b—variance-manipulated monosyllable training, Exp.2a—disyllable minimal pair training and Exp. 2b—disyllable non-minimal pair training)), using d' as DV. The results showed a significant main effect of Test ($F(1, 170)=7.5$, $p<.05$), a significant main effect of Tone pair ($F(5, 170)=6.8$, $p<.01$) and a significant Test x Experiment x Tone pair interaction ($F(10,170)=4.7$, $p<.05$). The main effects of Test and Tone pair are illustrated in Fig. 31.

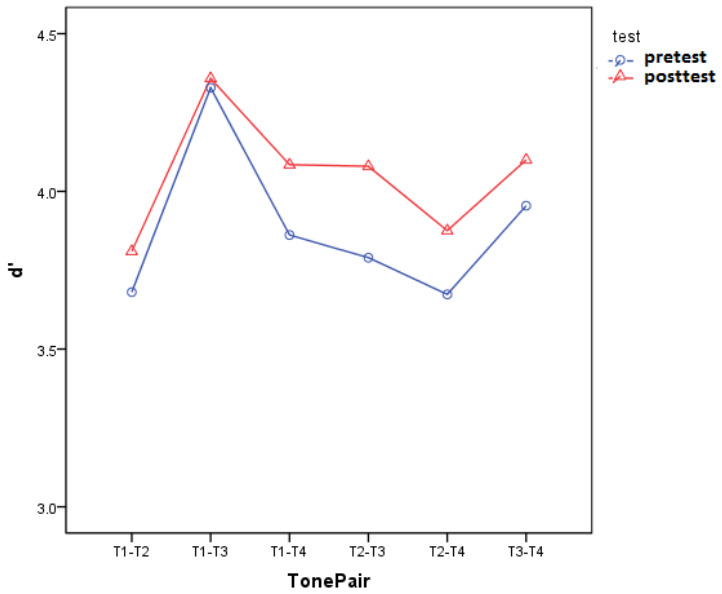

Figure 31. The d' averaged across the non-native control, variance manipulated and two disyllable training conditions in the pre- and posttest as a function of Tone pair.

As Fig. 31 shows, overall the d' increased after the training. The participants had the lowest d' for T1-T2 and T2-T4 whereas they had the highest d' for T1-T3. In terms of

discriminating T2-T4, the disyllable minimal pair training group had a very small d' increase whereas the disyllable non-minimal pair training group even had a d' decrease as shown in Fig. 30 (d) and 30 (e) respectively. Since the disyllable training conditions did not make participants shift their cue-weighting towards pitch direction, the lack of d' increase for T2-T4 among the two disyllable training groups again seemed to be related to the lack of cue-weighting shift towards pitch direction.

4.3.3 Sensitivity to lexical tones in the disyllable context

Same as the sensitivity analysis conducted for the tone discrimination performance in monosyllables, we first conducted a 2x3 mixed ANOVA (Within-subject: Test (pretest vs. posttest); Between-subject: Experiment (Exp. 1a—non-native control, Exp. 1b—variance-manipulated monosyllable training and Exp.1c—multi-talker monosyllable training)) using d' as DV to examine whether participants' sensitivity to lexical tones in disyllables improved after the monosyllable training. The result showed a significant main effect of Test ($F(1,27)=10.28$, $p<.01$) and a significant Test x Training interaction ($F(2,27)=2.9$, $p<.05$). The result is shown in Fig. 32.

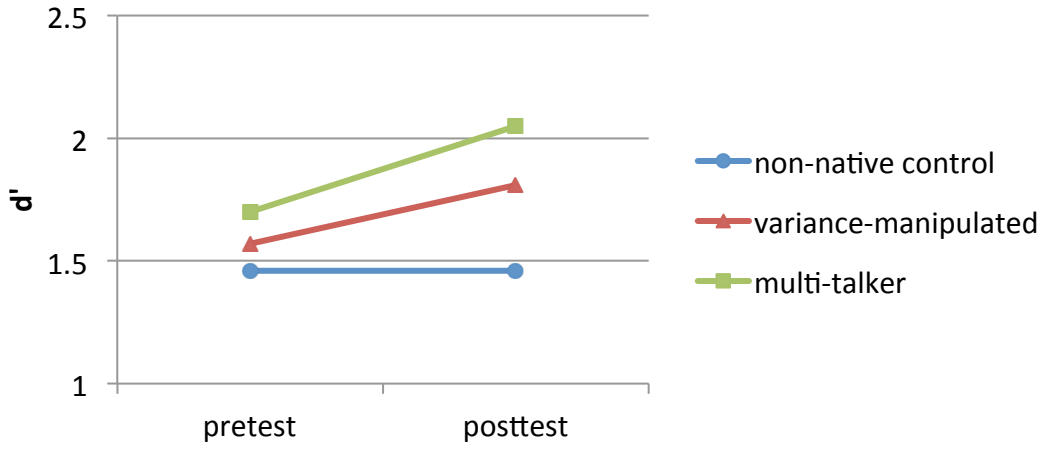

Figure 32. d' scores of the non-native control group and the two monosyllable training groups in the disyllable context before and after the training.

As shown in Fig. 32, the multi-talker training group had a larger d' increase than the variance-manipulated training group whereas the non-native control group had no d' change. Fig. 33 shows the simple effects of Test on the three groups.

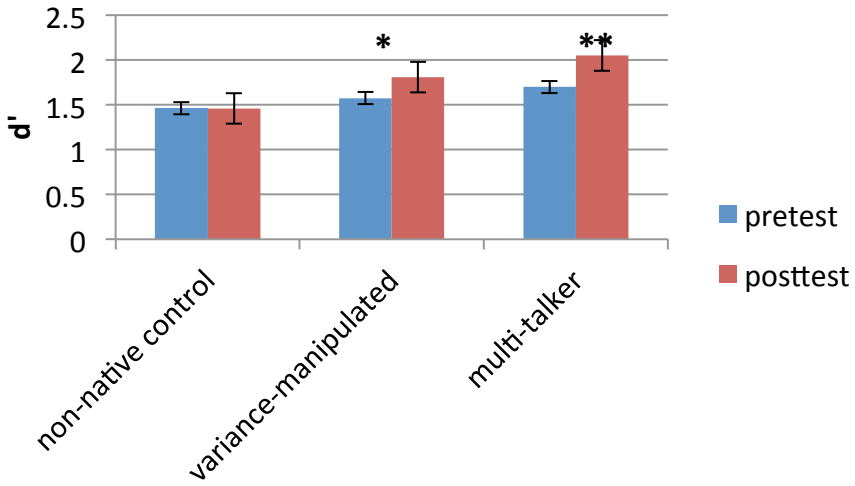

Figure 33. Simple effects of Test on the d' scores of the non-native control group and the two monosyllable training groups in the disyllable context before and after the training. *<.05, **<.01

As shown in Fig. 33, the non-native control group did not have a d' increase for tone discrimination in the disyllable context after the training. Thus, it suggested that there was no practice effect for tone discrimination in the disyllable context. The absence of practice effect on the tone discrimination in the disyllable context was likely due to a higher demand for tone categorization in the disyllable context (e.g., factoring out coarticulation effect).

We then conducted another 2x4 repeated measures ANOVA (Within-subject: Test (pretest vs. posttest); Between-subject: Experiments (Exp. 1a—non-native control, Exp. 1b—variance-manipulated monosyllable training, Exp. 2a—minimal pair disyllable training that used variance-manipulated stimuli and Exp. 2b—non-minimal disyllable pair training that used variance-manipulated stimuli) using d' as DV to examine whether disyllable training improved sensitivity to tones in the disyllable context. The results showed a significant main effect of Test ($F(1,36)=17.2$, $p<.01$) and a significant Test x Training interaction ($F(3,36)=6.6$, $p<.05$). Fig. 34 illustrates the Test x Training interaction.

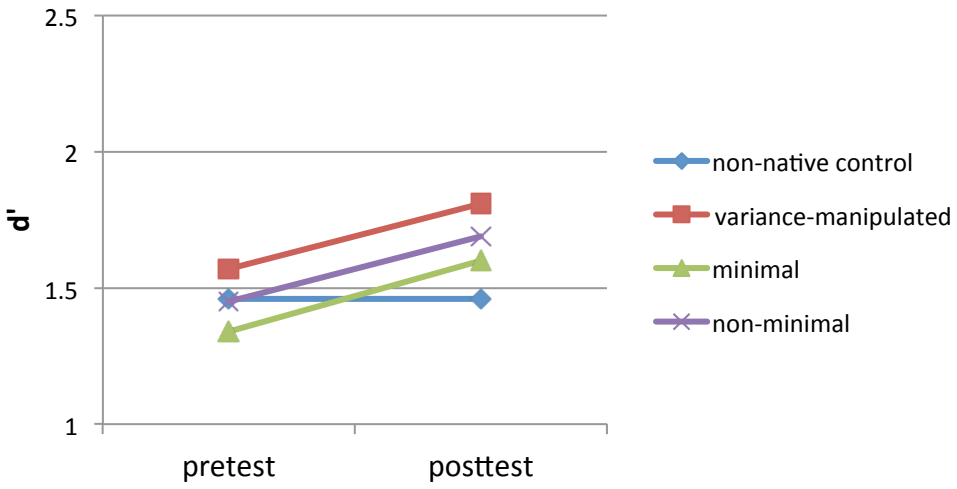

Figure 34. d' scores of the variance-manipulated group and the two disyllable training groups in the disyllable context before and after the training.

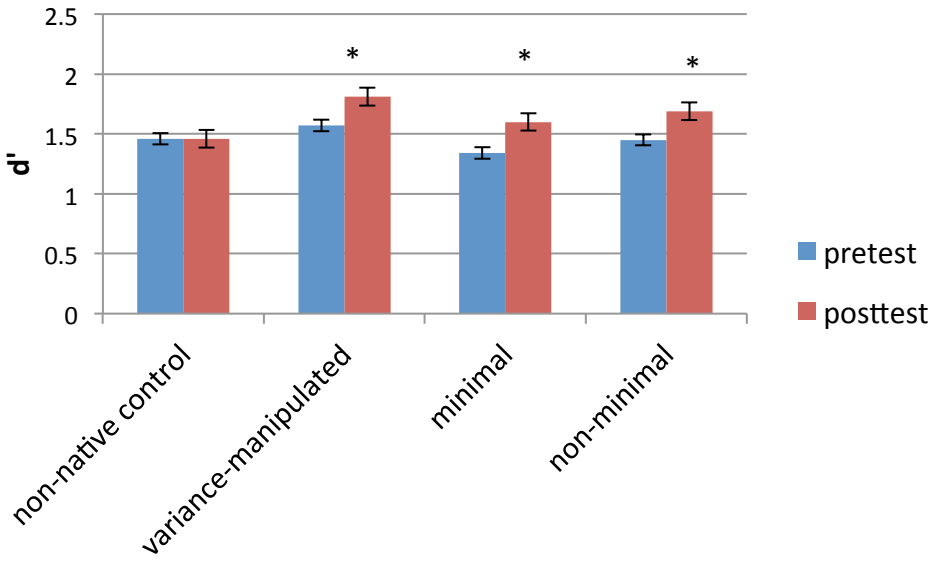

Figure 35. Simple effects of Test on the d' scores of the non-native control, the variance manipulated training and the two disyllable training groups before and after the training. *<.05.

As Fig. 34 shows, participants in the variance manipulated training, disyllable minimal training and disyllable non-minimal training consistently had d' increase for the

tone discrimination in disyllables after the training. The simple effect analysis showed that all three training groups except the non-native control group had a d' increase after the training, as shown in Fig. 35.

4.3.4 Tone discrimination for specific tone pairs in disyllable context

Similar to calculating d' for specific tone pairs in the monosyllable context, we calculated d' for specific tone pairs in the disyllable context as well. We conducted repeated measures ANOVA to examine the d' for six different tone pairs for the monosyllable training groups and the disyllable training groups respectively. First, we conducted a 2x6x3 repeated measures ANOVA to examine the non-native control and two monosyllable training groups' sensitivity for discriminating specific tone pairs (Within-subject: Test (pretest vs. posttest); Tone pair (T1 vs. T2, T1 vs. T3, T1 vs. T4, T2 vs. T3, T2 vs. T4 and T3 vs. T4); Between-subject: Experiment (Exp. 1a—monosyllables without tone training/ non-native control, Exp. 1b—variance-manipulated monosyllable training and Exp.1c—multi-talker monosyllable training)), using d' as DV. The results showed a significant main effect of Test ($F(1, 125)=13.2$, $p<.01$), a significant main effect of Tone pair ($F(5, 125)=7.8$, $p<.01$) and a significant Test x Experiment ($F(3,125)=4.7$, $p<.05$). The significant Test x Experiment interaction was very similar to the one we reported Section 4.3.1, namely, the non-native control group did not have d' increase after the training whereas the two monosyllable training groups had d' increase. Thus, here we only illustrate the main effects of Test and Tone pair in Fig. 36.

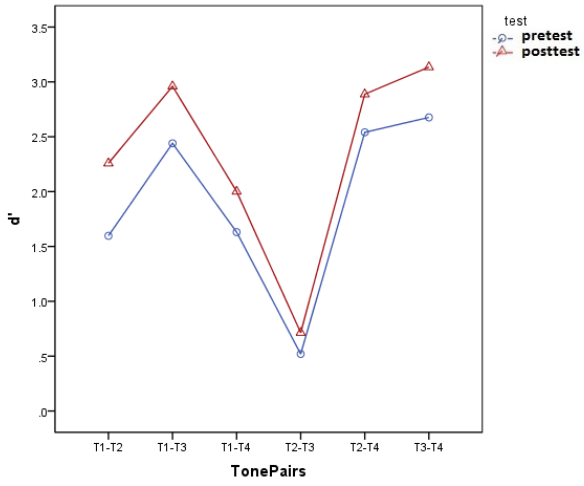

Figure 36. The d' averaged across the non-native control and two monosyllable training conditions in the pre- and posttest as a function of Tone pair

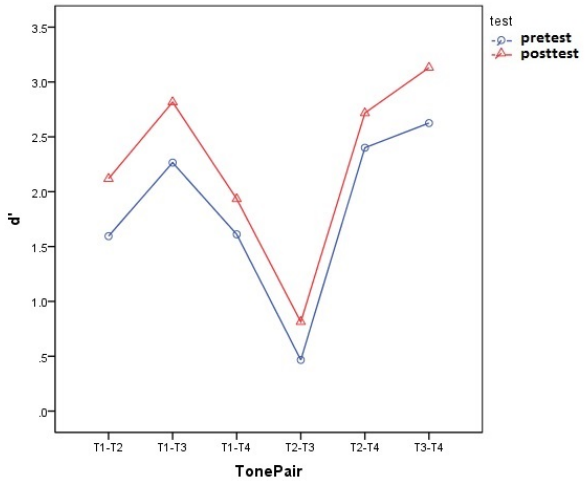

Figure 37. The d' averaged across the non-native control, variance-manipulated training and two disyllable training conditions in the pre- and posttest as a function of Tone pair

As Fig. 36(a) shows, the overall d' increased for all tone pairs in the posttest relative to the pretest. In both pretest and posttest, the discrimination between T2 and T3 was the poorest. Bear in mind that all target test stimuli in the non-speeded AX discrimination task were preceded by T1, T2 or T4. The preceding tones, T1, a high level tone and T2, a rising tone gave T2 in the second syllable a clear initial fall before the rise. Such tonal coarticulation made the pitch contour of T2 similar to that of T3, which also had an initial falling followed by a final rising pitch. On the other hand, the preceding tones somehow made the discrimination between T2 and T4 on the second syllable easier. However, the d' for T2-T4 discrimination in the disyllable context was still lower than that in the monosyllable context. As for why the disyllable training conditions did not outperform the monosyllable training in the tone discrimination task in the disyllable context, it may be due to the lack of natural coarticulation in the resynthesized disyllables

(Chapter Five has more discussion on the result of sensitivity difference among specific tone pairs in the disyllable context).

We conducted another 2x6x4 repeated measures ANOVA to examine the participants' sensitivity to the difference of specific tone pairs for the disyllable training groups (Within-subject: Test (pretest vs. posttest); Tone pair (T1 vs. T2, T1 vs. T3, T1 vs. T4, T2 vs. T3, T2 vs. T4 and T3 vs. T4); Between-subject: Experiment (Exp. 1a—monosyllables without tone training/ non-native control, Exp. 1b—variance-manipulated monosyllable training, Exp.2a—disyllable minimal pair training and Exp. 2b—disyllable non-minimal pair training)), using d' as DV. The results showed a significant main effect of Test ($F(1, 170)=4.4$, $p<.05$), a significant main effect of Tone pair ($F(5, 170)=8.8$, $p<.01$) and a significant Test x Experiment interaction ($F(4,170)=3.6$, $p<.05$). Again, the significant Test x Experiment interaction was very similar to the one we reported in Section 4.3.3, namely, the non-native control group did not have a d' increase after the training whereas the variance-manipulated training group and the two disyllable training groups did. Thus, here we only illustrate the main effects of Test and Tone pair in Fig. 37. Overall, the pattern was very similar to the one found for the monosyllable training groups' tone discrimination in the disyllable context, namely, the discrimination for T2-T3 was the worst.

4.3.5 Summary of sensitivity to lexical tones in monosyllables and disyllables

Naïve listeners (Native English speakers) in all groups including the non-native control group showed a significant d' increase for tone discrimination in monosyllables

after the video-game training. We argue that the d' increase in the non-native control group may have come from the feedback provided after each trial, causing a practice effect. It also suggests that in general native English speakers are sensitive to tone differences in the monosyllable context. However, for tone discrimination in the disyllable context, naïve listeners performed significantly worse than in monosyllables in the pretest. After the training, all trainee groups had a significant d' increase whereas the non-native control group did not show any d' increase for tone discrimination in disyllables. Since there was no feedback in the test, the d' increase must be due to the training condition. Regardless of the training condition, the performance of tone discrimination in disyllables was still worse than that in monosyllables after the training. In terms of the d' improvement for tone discrimination in disyllables, among the monosyllable training groups, the multi-talker training group had a larger d' improvement than variance-manipulated training group whereas among the disyllable training groups, both the minimal pair and non-minimal pair groups had the same amount of improvement as the variance-manipulated training group. In terms of the discrimination for specific tone pairs in the monosyllable context, T1-T3 was discriminated better than T2-T4 in the pretest for all groups. After the training, the variance-manipulated and multi-talker training groups showed a d' increase for T2-T4 whereas other training groups and the non-native control group did not. These results suggest that the discrimination for specific tone pairs is related to the cue-weighting on pitch height and pitch direction as only the participants in the variance-manipulated training and multi-talker training conditions shifted their cue-weighting towards pitch direction and their perceptual

distance between T2 and T4 increased. Thus, only the variance-manipulated training and multi-talker training groups' discrimination for T2-T4 improved. In terms of the discrimination of specific tone pairs in the disyllable context, all participants had more difficulty discriminating T2-T3 relative to other tone pairs. We argue that the difficulty discriminating T2-T3 came from the coarticulation effect, as T1 and T2 preceded the target tone stimuli in the disyllable context, making the pitch contour of T2 and T3 similar to each other. It seemed that the shift of cue-weighting towards pitch direction did not help the tone discrimination in the disyllable context.

4.4 Word identification results

The non-native control group and non-minimal disyllable training group reached ceiling in word identification for both the tone tokens used in the training (mean accuracy rate: 100%) and the tone tokens produced by new talkers (mean accuracy rate: 100%). The ceiling effect in these two groups is not surprising because the participants in both groups completely relied on the four syllables contrasted by segments for the association with the animals (*ma, na, sa* and *fa* for the control group; *ta1*, *ku1*, *po1*, *ti1* for the non-minimal disyllable training group). To confirm that the participants were using the first syllables in the non-minimal pair disyllable training condition, at the end of the word identification task, we asked the participants to tell whether they were using the first syllable or the second syllable for playing the video game. They consistently reported that they were using the first syllables to play the game. Focusing only on the first syllables

but not the second syllables that carried contrastive tones may explain why the non-minimal disyllable pair training condition did not improve the tone discrimination performance as much as we expected.

For word identification, we were primarily concerned with whether participants in the variance-manipulated training (Exp. 1b), multi-talker training (Exp. 1c) and disyllable minimal pair training (Exp. 2a) differed in terms of accuracy rates because tones were contrastive in these three training conditions. All three training groups' word identification accuracy rates were above chance level for both the old talker and the new talker stimuli. Especially for the word identification of the stimuli used in the training, the accuracy rates were well above the chance level (variance-manipulated training group: 76%; multi-talker training group: 74%; minimal pair disyllable training group: 88%), suggesting that the trainees in the three training groups associated the four lexical tones with the animals after playing the video game. To study the three training groups' word identification performance quantitatively, we conducted a mixed effect logistic regression using three categorical predictors (Talker—old stimuli vs. new stimuli; Tone—4 lexical tones; Training—(1) Exp. 1b: variance manipulated training, (2) Exp. 1c: multi-talker training, (3) Exp.2b: minimal pair disyllable training) and subject as a random effect to predict the word identification accuracy rate, which was transformed into logit[3] (the higher log odds, the higher probability of making correct responses). The logistic

[3] Since the outcome of the word identification task is dichotomous (i.e., correct or incorrect), according to Baayen (2008, pp195-202), it is more accurate to use logit instead of the proportion of correct responses (i.e., accuracy rate) as the dependent variable in logistic regression. The logit was calculated as the logarithm of odds ratio between the correct responses to the incorrect responses.

logit = log(nCorrect/nIncorrect) where the base is natural logarithm.

regression allows us to examine, first, whether word/tone identification is different between the old talker stimuli and new talker stimuli; second, whether the word identification accuracy rates for the four different lexical tones differed; and third, whether the three training groups differed in terms of word identification accuracy rates. In order to make all these comparisons, we repeated the logistic regressions by changing the baselines of Talker, Training and Tone. We ran the mixed effect logistic regressions sequentially, first without any interaction term, then with all 2-way interactions, finally with the 3-way interaction, using the lmer() function in lme4 package in R.

First, we compared the three models with and without interactions. The maximum likelihood comparisons showed that adding all 2-way interactions significantly improved the model that did not include interaction (χ^2=111, *df*=11, p<.001) and adding the 3-way interaction improved the model that included all 2-way interactions with marginal significance (χ^2=17, *df*=6, p=.07). Thus, we repeated the mixed effect logistic regressions with all 2-way interactions that used different categories as baselines. The first model is summarized in the following table.

Table7. Logistic Regression Analysis of three training groups' word identification accuracy rate, using old talker stimuli, T4 and multi-talker training condition as the baselines.

	β	Std. Error	z value	P (2-tailed)	
(Intercept)	0.96122	0.35159	2.734	0.006259	**
Talkernew	0.57882	0.38215	1.515	0.129864	
Tone1	-0.18844	0.35186	-0.536	0.59228	
Tone2	0.06538	0.35889	0.182	0.855451	
Tone3	1.42753	0.45622	3.129	0.001754	**
Trainingvm	-0.21327	0.49352	-0.432	0.665642	

Trainingminimal	-0.04042	0.49873	-0.081	0.935409	
Talkernew:Tone1	0.02917	0.53093	0.055	0.956181	
Talkernew:Tone2	-0.37268	0.53093	-0.702	0.482717	
Talkernew:Tone3	-1.2536	0.61742	-2.03	0.042319	*
Talkernew:Trainingvm	-1.21533	0.5117	-2.375	0.017545	*
Talkernew:Trainingminimal	-1.3739	0.51899	-2.647	0.008115	**
Tone1:Trainingvm	0.88102	0.51451	1.712	0.086834	.
Tone2:Trainingvm	0.79519	0.52664	1.51	0.131064	
Tone3:Trainingvm	-1.30217	0.57615	-2.26	0.023813	*
Tone1:Trainingminimal	2.81705	0.74397	3.786	0.000153	***
Tone2:Trainingminimal	1.63192	0.60121	2.714	0.00664	**
Tone3:Trainingminimal	1.20109	0.79864	1.504	0.132603	

N=240, '***' <0.001 '**' <0.01 '*' <0.05 '.' marginal significance.

The intercept in Table 7 was interpreted as the word identification log odds of the multi-talker training group for T4 for the old talker stimuli. In terms of the talker effect, the multi-talker training group's identification of T4 did not differ between the old talker stimuli and new talker stimuli (Talkernew: β=0.57882, z=1.515, p=0.129864). The result showed a significant Talker by Tone interaction (Talkernew:Tone3: β=-1.2536, z=-2.03, p=0.042319). The significant interaction suggested that the log odds of the multi-talker training group's identification of T3 for the new talker stimuli was significantly lower than that for the old talker stimuli. The multi-talker training group's identification of T1, T2 and T4 in the old talker stimuli was the same as the one in the new talker stimuli. The Talker by Tone interaction for the multi-talker training group was illustrated in Fig. 38:

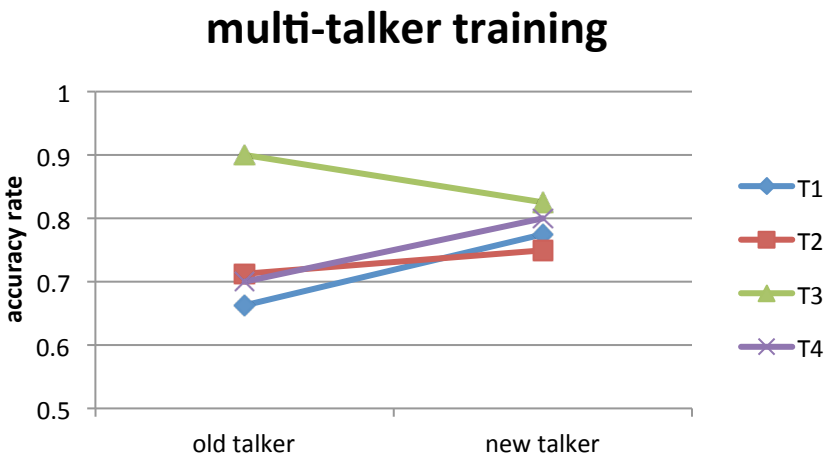

Figure 38. The multi-talker training group's word identification accuracy rates of different tones.

In terms of the training effect, the result showed the variance-manipulated training group and the minimal pair disyllable training group did not differ from the multi-talker training group in terms of log odds for T4 in the old talker stimuli (Trainingvm: β=-0.21327, z=-0.432, p=0.665642; Trainingminimal: β=-0.0404, z=-0.081, p=0.935409). It suggested that there was no difference between the multi-talker training group and the other two groups in terms of the log odds for T4 in the old talker stimuli. There were significant Talker by Training interactions (Talkernew:Trainingvm: β= -1.21533, z=-2.375, p=0.017545; Talkernew:Trainingminimal: β= -1.3739, z= -2.647, p=0.008115). The interactions suggested that the log odds of the variance-manipulated training group and the minimal pair disyllable training group's identification of T4 in the old talker stimuli were significantly higher than that in the new talker stimuli. The Talker by Training interaction for T4 is illustrated in Fig. 39d.

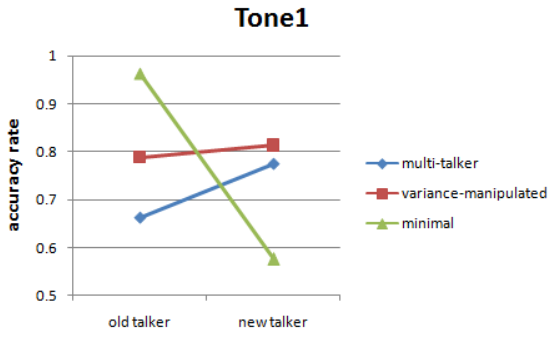

Figure 39a

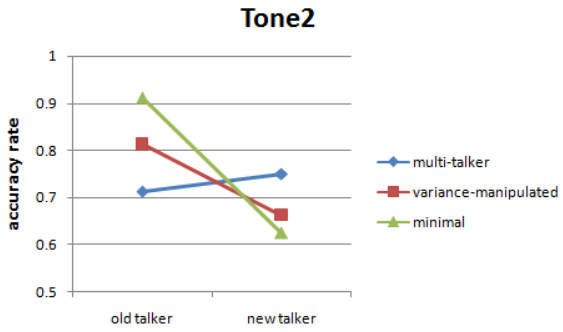

Figure 39b

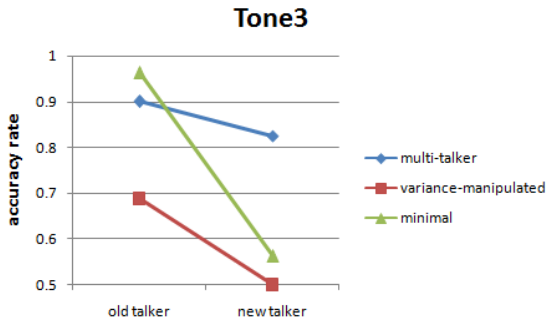

Figure 39c

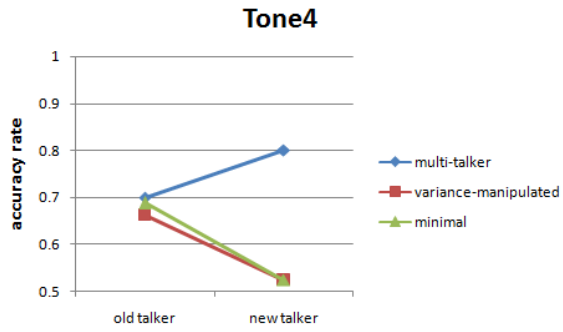

Figure 39d

Figure 39. Three training groups' word identification accuracy rates for the old and new talker stimuli for each lexical tone.

We further examined the Talker by Training interaction for T3, T2 and T1 by changing the baseline of the Tone (see Tables A, B and C in Appendix). As Fig. 39c showed, the minimal pair disyllable training group's log odds for T3 in the new talker stimuli was significantly lower than that for T3 in the old talker stimuli (Talkernew: Trainingminimal: β=-2.5726, z= -3.169, p=0.001531, see Table A in Appendix) whereas the multi-talker training group and the variance-manipulated training group had the same log odds for T3 in the old talker stimuli and the new talker stimuli. As Fig. 39b shows, for T2, both the variance-manipulated training group and the minimal pair disyllable

training group's log odds for the new talker stimuli were significantly lower than that for the old talker stimuli (Talkernew:Trainingvm: β=-1.0667, z=-2, p=0.0455; Talkernew:Trainingminimal: β=-2.22153, z=-3.68, p=0.000233, see Table B in Appendix). The multi-talker training group and the variance-manipulated training group had the same log odds for T2 in the old talker stimuli and in the new talker stimuli. As Fig. 39a shows, for T1, the minimal pair disyllable training group's log odds for the new talker stimuli was significantly lower than that for the new talker stimuli (Talkernew:Trainingminimal: β=-3.79612, z=-5.072, p<.001, see Table C in Appendix). The multi-talker training group and the variance-manipulated training group had the same log odds for the old talker stimuli and the new talker stimuli for T1.

There were also significant Tone by Training interactions (Tone3: Trainingvm: β=-1.30217, z=-2.26, p=0.023813; Tone1:Trainingminimal: β=2.81705, z=3.786, p=0.000153; Tone2:Trainingminimal: β=1.63192, z=2.714, p=0.00664). The Tone by Training interactions suggested that, for the old talker stimuli, the variance-manipulated training group's log odds for T3 was significantly lower than the multi-talker training group's log odds for T3. For T1 and T2 in the old talker stimuli, the minimal pair disyllable training group's log odds was significantly higher than the multi-talker training group. The Tone by Training interaction for the old talker stimuli is illustrated in Fig. 40a.

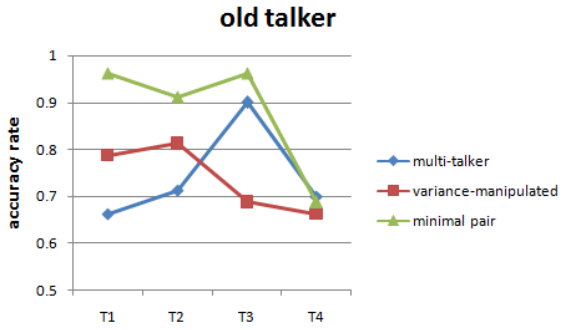

Figure 40a

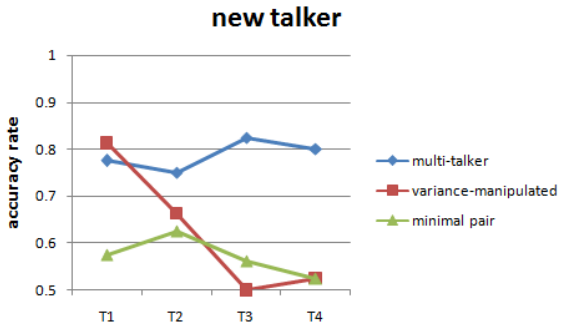

Figure 40b

Figure 40. Three training groups' word identification accuracy rates for different tones in the old talker stimuli and new talker stimuli.

We further examined the Tone x Training interaction for the new talker stimuli by changing the baseline of Talker to be new talker stimuli (see Table D in Appendix). The results showed the multi-talker training group's log odds was significantly higher than the variance-manipulated training group's log odds (Trainingvm: β=-1.4286, z=-2.82, p=0.0048) and the minimal pair disyllable training group's log odds (Trainingminimal: β=-1.41432, z=-2.777, p=0.00548). Also, the interactions showed that both the variance-manipulated training group's log odds and the minimal pair disyllable training group's log odds were significantly lower than the multi-talker training group's log odds (Tone3:Trainingvm: β=-2.28527, z=-2.536, p=0.00161; Tone3:Trainingminimal: β=-2.17375, z=-3.004, p=0.00647). There was also another interaction (Tone1:Trainingminimal: β=-3.34826, z=-2.754, p=0.0006). The Tone x Training interactions for new talker stimuli are illustrated in Fig. 40b. As Fig. 40b shows, the multi-talker training group outperformed the other two groups for T3 and T4 in the new

talker stimuli. Also, the multi-talker training group identified T1 in the new talker stimuli better than the minimal pair training group.

In order to examine the Talker by Tone interaction for the variance manipulated training group, we changed the baseline of Training to be the variance manipulated training condition. The result is shown in Table 8. When the baseline of Tone was T4, the variance manipulated training group's log odds for the new talker stimuli was significantly lower than that for the old talker stimuli (Talkernew: β=-0.6365, z=-1.87, p=0.041417). There was a significant Talker x Tone interaction (Talkernew:Tone3: β=-1.2367, z=-2.559, p=0.004151). The interaction suggested that the variance-manipulated training group's log odds for T3 in the new talker stimuli was significantly lower than that in the old talker stimuli. Thus, the variance-manipulated training group's identification of T4 and T3 was worse in the new talker stimuli. The Talker x Tone interaction for the variance-manipulated training group is illustrated in Fig. 41.

Table8. Logistic Regression Analysis of three training groups' word identification accuracy rate, using old talker stimuli, T4 and variance manipulated training condition (vm) as the baselines.

	β	Std. Error	z value	P (2-tailed)	
(Intercept)	0.7479	0.3463	2.16	0.030797	*
Talkernew	-0.6365	0.3403	-1.87	0.041417	*
Tone1	0.6926	0.3754	1.845	0.065042	.
Tone2	0.8606	0.3854	2.233	0.025562	*
Tone3	0.1254	0.3519	0.356	0.721655	
Trainingmultitalker	0.2133	0.4935	0.432	0.665602	
Trainingminimal	0.1728	0.495	0.349	0.726964	
Talkernew:Tone1	0.8045	0.5324	1.511	0.130742	
Talkernew:Tone2	-0.2241	0.5137	-0.436	0.662713	

Talkernew:Tone3	-1.2367	0.4831	-2.559	0.004151	**
Talkernew:Trainingmultitalker	1.2153	0.5117	2.075	0.057546	.
Talkernew:Trainingminimal	-2.1586	0.489	-2.324	0.005727	**
Tone1:Trainingmultitalker	-0.881	0.5145	-1.712	0.086835	.
Tone2:Trainingmultitalker	-0.7952	0.5266	-1.51	0.131053	
Tone3:Trainingmultitalker	1.3022	0.5761	2.26	0.023813	*
Tone1:Trainingminimal	1.936	0.7554	2.563	0.010378	*
Tone2:Trainingminimal	0.8367	0.6174	1.355	0.175353	
Tone3:Trainingminimal	2.5033	0.744	3.365	0.000766	***

N=240, '***' <0.001 '**' <0.01 '*' <0.05 '.' marginal significance.

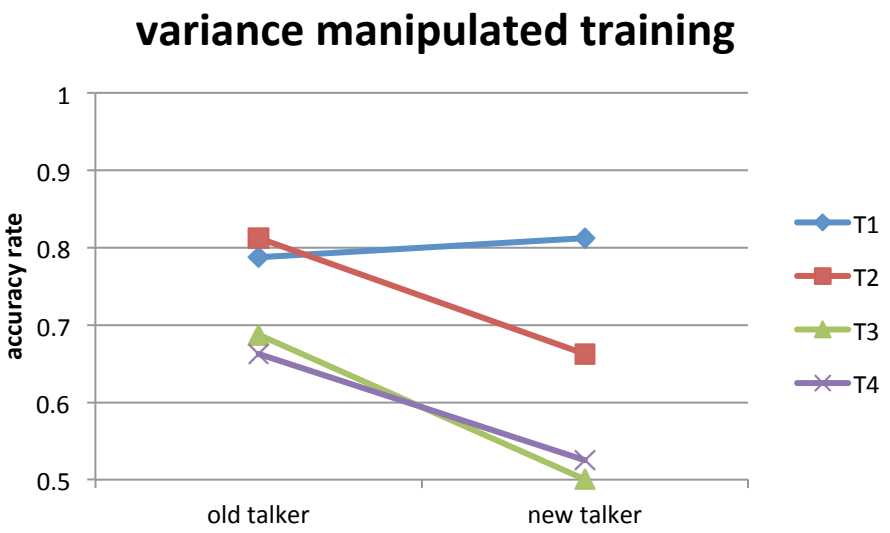

Figure 41. Word/Tone identification of the variance manipulated training group for the old talker and new talker stimuli.

In Table 8, there were also Tone x Training interactions for the old talker stimuli. The interaction term Tone3: Trainingmultitalker (β=1.3022, z=2.26, p=0.023813) suggested that, for the old talker stimuli, the log odds of the multi-talker training group for T3 was significantly higher than that of the variance manipulated training group. The results also showed that the minimal pair disyllable training group's log odds was higher than the variance-manipulated training group for T1 and T3 (Tone1:Trainingminimal:

β=1.936, z=2.563, p=0.010378; Tone3:Trainingminimal: β=2.5033, z=3.365, p=0.000766). The Tone x Training interaction effect in the old talker stimuli is illustrated in Fig. 40a. We also changed the baseline of Talker to be new talker stimuli for the model where the variance-manipulated training group was the baseline (see Table E in Appendix). The results showed that the log odds of the multi-talker training group was significantly higher than the variance-manipulated training group, which was the result we found in the model where the multi-talker training condition was set as the baseline (Trainingmultitalker: β=1.4286, z=2.82, p=0.004795). For the new talker stimuli, there was also a Tone by Training interaction effect, which is shown in Fig. 40b. The results showed that, for T3, the log odds of the multi-talker training group was significantly higher than that of the variance-manipulated training group (Tone3:Trainingmultitalker: β=2.28528, z=4.297, p=0.001607) and there was no difference between the variance-manipulated training group and the minimal pair disyllable training group. For T1, the log odds of the minimal pair disyllable training group was significantly lower than that of the variance manipulated training group (Tone1:Trainingminimal: β=-1.26152, z=-2.478, p=0.013228). Together with the results of the Tone x Training interactions found in the models where the multi-talker training group was the baseline, the Tone x Training interaction results can be summarized as follows: In the old talker stimuli, the minimal pair disyllable training group identified T1 and T2 the best whereas the minimal pair disyllable training group and the multi-talker training group identified T3 better than the variance-manipulated training group. For T4, there was no difference among the three training groups. In the new talker stimuli, the multi-talker training group and the

variance-manipulated training group identified T1 better than the minimal pair disyllable training group. The multi-talker training group identified T3 and T4 better than the variance-manipulated training group and the minimal pair disyllable training group. For T2, there was no difference among the three training groups.

Finally, we ran a model with the minimal pair disyllable training group as the baseline in order to examine the Talker x Tone interaction for this particular training group. The result is shown in Table 9. The result showed that the log odds of the minimal pair disyllable training group's for T4 in the old talker was significantly higher than that in the new talker stimuli (Talkernew: β=-0.79507, z=-2.264, p=0.023567). The Talker x Tone interaction showed that T1, T2 and T3 in the old talker stimuli had significantly lower log odds than those in the new talker stimuli (Talkernew:Tone1: β=-2.39306, z=-3.241, p=0.00119; Talkernew:Tone2: β=-1.22031, z=-2.063, p=0.039135; Talkernew:Tone3: β=-2.45232, z=-3.322, p=0.000893). The Talker by Tone interaction for the minimal pair disyllable training group is illustrated in Fig. 42.

Table 9. Logistic Regression Analysis of three training groups' word identification accuracy rate, using old talker stimuli, T4 and minimal pair disyllable training (minimal) condition as the baselines.

	β	Std. Error	z value	P (2-tailed)	
(Intercept)	0.9208	0.35372	2.603	0.009236	**
Talkernew	-0.79507	0.35116	-2.264	0.023567	*
Tone1	2.62861	0.6555	4.01	6.07E-05	***
Tone2	1.6973	0.48234	3.519	0.000433	***
Tone3	2.62861	0.6555	4.01	6.07E-05	***
Trainingvm	-0.17286	0.49503	-0.349	0.72695	
Trainingmultitalker	0.04043	0.49873	0.081	0.93539	

Talkernew:Tone1	-2.39306	0.7383	-3.241	0.00119	**
Talkernew:Tone2	-1.22031	0.59159	-2.063	0.039135	*
Talkernew:Tone3	-2.45232	0.73815	-3.322	0.000893	***
Talkernew:Trainingvm	0.15857	0.48899	0.324	0.745721	
Talkernew:Trainingmultitalker	1.3739	0.51899	2.647	0.008115	**
Tone1:Trainingvm	-1.93602	0.75538	-2.563	0.010378	*
Tone2:Trainingvm	-0.83672	0.61741	-1.355	0.175352	
Tone3:Trainingvm	-2.50326	0.74397	-3.365	0.000766	***
Tone1:Trainingmultitalker	-2.81703	0.74397	-3.786	0.000153	***
Tone2:Trainingmultitalker	-1.63193	0.60121	-2.714	0.006639	**
Tone3:Trainingmultitalker	-1.2011	0.79864	-1.504	0.1326	

N=240, '***' <0.001 '**' <0.01 '*' <0.05 '.' marginal significance.

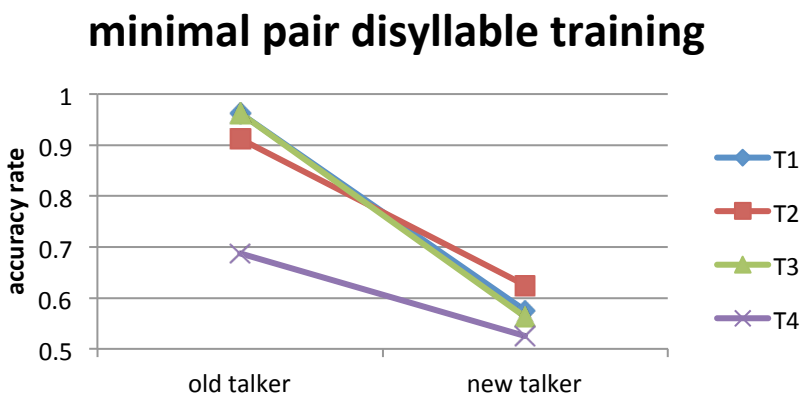

Figure 42. Word/Tone identification of the minimal pair disyllable training group for the old talker and new talker stimuli.

4.4.1 Summary of word identification results

With the logistic regression analysis of the word identification results, we provided a detailed description of the interactions between different predictors (Talker, Tone and Training). The non-native control group and the four training groups demonstrated that they all associated the words that occurred in the video game with the four different animals. The results suggested that the implicit word learning training paradigm had successfully allowed the learners to make sound meaning associations.

Within the three groups trained either on monosyllables or disyllables with contrastive tones, the result showed that the multi-talker training group outperformed the variance-manipulated training group and the minimal pair disyllable training group in terms of generalization to new talkers for T2, T3 and T4. For T1, the multi-talker training group did not outperform the variance manipulated training group. Overall, the multi-talker training was more robust for making tone identification generalizations to the new talker stimuli. In terms of the identification of particular lexical tones, in the old talker stimuli, the minimal pair disyllable training group outperformed the other two training groups for T1 and T2; the minimal pair disyllable training and the multi-talker training group identified T3 better than the variance manipulated training group; there was no difference among the three training groups for the identification of T4. In the new talker stimuli, the multi-talker training group and the variance-manipulated training group identified T1 better than the minimal pair disyllable training group. The multi-talker training group identified T3 and T4 better than the variance manipulated training group and the minimal pair disyllable training group. For T2, there was no difference among the three training groups. Since the identification of the new talker stimuli is an important indicator of learning, the multi-talker training seemed to be the most robust, particularly for the learning of T3 and T4.

4.5 Relation between game performance and tone categorization

There are several ways to examine the participants' game performance. To examine the game performance visually, we plotted the participants' word identification accuracy rates in the game (game accuracy rate henceforth) from Day 1 to Day 4. We found that regardless of the training condition, there was a clear pattern of individual differences in terms of the video-game playing. Figs. 43, 44 and 45 illustrate the word identification accuracy rates through Day 1 to Day4 of four participants randomly selected from the variance manipulated training, the multi-talker training and the minimal pair disyllable training group as examples to demonstrate the individual differences.[4] Since the accuracy rate before level 5 was always 100% for all participants as the animals were visually clear before level 5. We only plotted the accuracy rates from level 5.

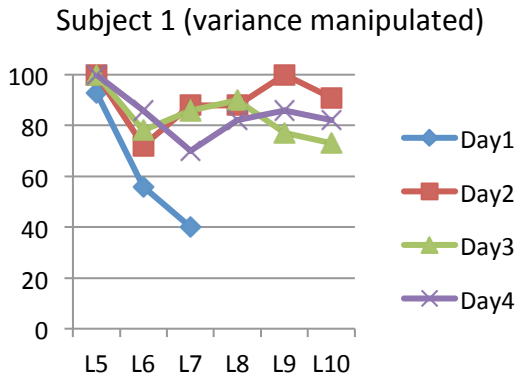

Figure 43 (a)

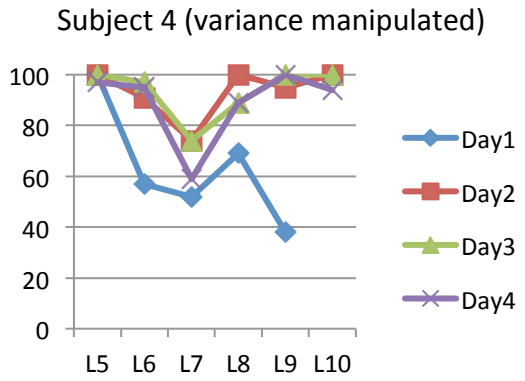

Figure 43 (b)

[4] The reason we did not report the game performance of the control group and the non-minimal pair disyllable training group is that both group reached the ceiling soon during the four days' video game playing as they were using the segmental information rather than the tone information to play the game.

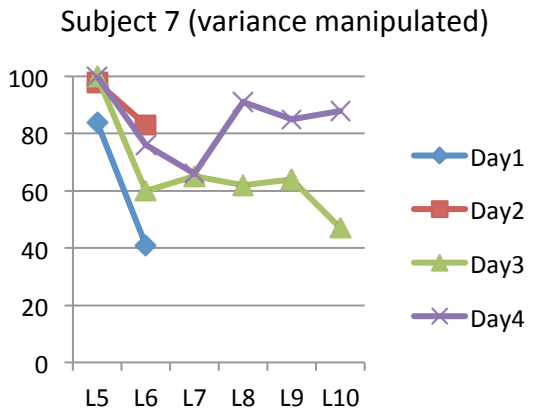

Figure 43 (c)

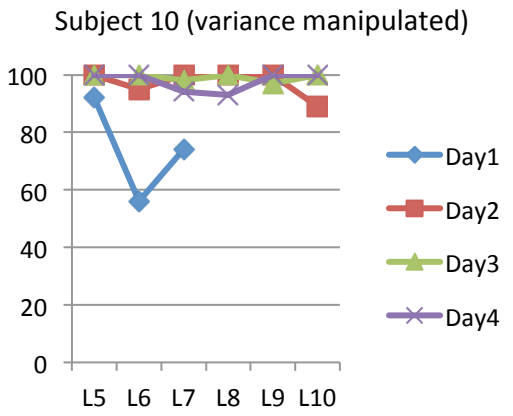

Figure 43 (d)

Figure 43. Four participants in the variance manipulated training group's word identification accuracy rates from level 5 to level 10 in four days.

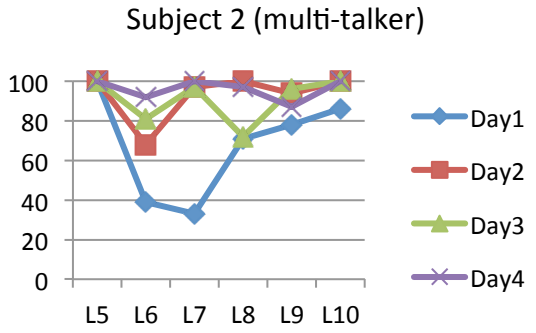

Figure 44 (a)

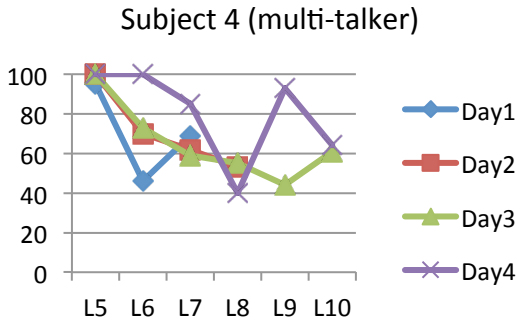

Figure 44 (b)

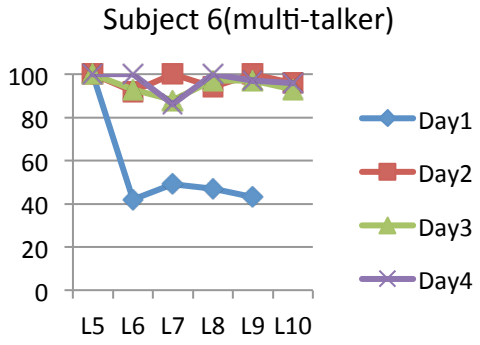

Figure 44 (c)

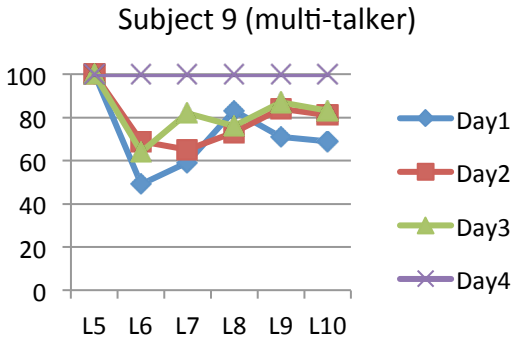

Figure 44 (d)

Figure 44. Four participants in the multi-talker training group's word identification accuracy rates from level 5 to level 10 in four days.

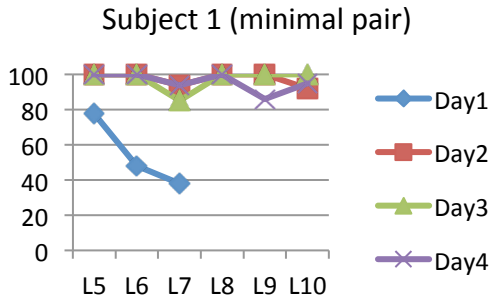

Figure 45 (a)

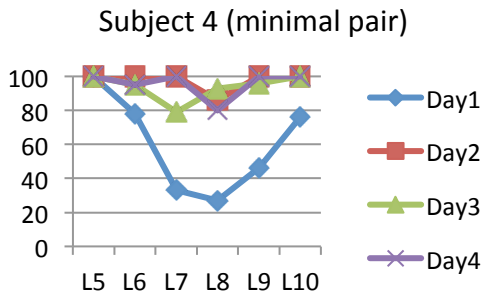

Figure 45 (b)

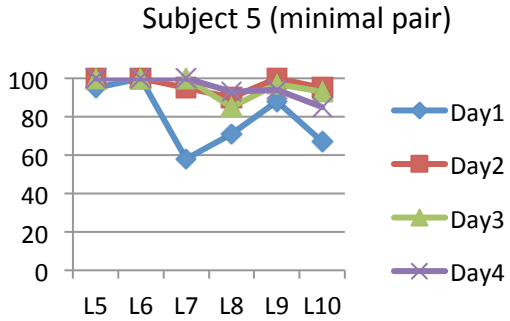

Figure 45 (c)

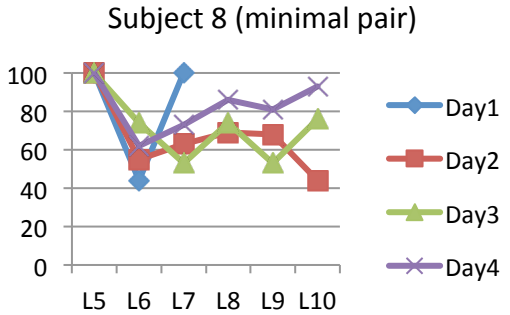

Figure45 (d)

Figure 45. Four participants in the minimal pair disyllable training group's word identification accuracy rates from level 5 to level 10 in four days.

As Fig. 43, Fig. 44 and Fig. 45 show, on Day 1 most of the participants had a much lower accuracy rate starting from level 7, which was the stage where the four animals could no longer be seen clearly and the speed of the animals' movement increased. In fact, none of the participants in the variance manipulated training group reached level 10 on Day 1 whereas a couple of participants in the multi-talker and minimal pair disyllable training group did. Regardless of the training condition, most of the participants reached level 10 since Day 2 except for a handful of them (e.g., subject 7 in the variance manipulated training group and subject 4 in the multi-talker training group). On Day 3 and Day 4, all participants in the three training groups reached level 10.

As shown in the graphs, subject 9 in the multi-talker training group had reached a relatively high accuracy rate on Day 1 and reached ceiling effect on Day 4, whereas subject 10 in the variance manipulated training group failed to reach level 10 on Day 1, but reached ceiling since Day 2. These were just several examples to illustrate that the rates to reach the ceiling were different for different individuals. But overall, the results showed a clear sound to meaning association after playing the video game as the game accuracy rate reached the ceiling.

There were three parameters that can be used to quantify the participants' game performance—(1) the word/tone identification accuracy rate averaged from level 5 to level 10 (game accuracy rate); (2) the total number of tone tokens played during the four days' video game playing; (3) the total number of times of clearing level 10. There were also three parameters that can be used to quantify the participants' tone categorization performance—(1) d' for tone discrimination in disyllable context;[5] (2) the accuracy rate of word/tone identification for the old talker's stimuli in the posttest; (3) the accuracy rate of word/tone identification for the new talker's stimuli in the posttest. In order to examine the relationship between the game performance and the tone categorization, we conducted three hierarchical regressions, using the three parameters of game performance as the predictors for each of the three parameters of tone categorization performance. The hierarchical regressions can inform us whether the parameters of game performance can

[5] We did not use the d' score for tone discrimination in monosyllable context as the parameter to quantify the tone categorization performance because the participants already reached ceiling for tone discrimination in the monosyllable context.

predict the tone categorization performance in terms of both tone discrimination and tone identification. The results are reported in the following two sections.

4.5.1 Relation between game performance and tone discrimination

At first, we ran a hierarchical regression that used three predictors—(1) the word/tone identification accuracy rate averaged from level 5 to level 10 (game accuracy rate); (2) the total number of tone tokens played during the four days' video-game playing; (3) the total number of times of clearing level 10 to predict the d' scores for tone discrimination in the disyllable context. We found a highly significant correlation between the game accuracy and the total number of times of clearing level 10 (r=0.76, n=30, p<.001). Thus, we used (1) the total number of tone tokens played during the four days' video-game playing; (2) the total number of times of clearing level 10 as the two predictors for d' in a hierarchical regression. We entered the two predictors in steps.[6] The result is summarized in the Table 10.

Table 10. Hierarchical regression result with two predictors: the total number of input tone tokens and the total number of times of clearing level 10 during the video game training. DV: d' scores for tone discrimination in the disyllable context.

	B	SE B	β (standardized B)	R^2	ΔR^2
Dependent variable:	d' scores for tone discrimination in the disyllable context				
Step 1				0.161*	0.161*
Total number of	0	0	0.401		

[6] Because there is no reference for us to decide the order of entering the predictors in the hierarchical regression, we entered the two predictors in both orders and the result was the same. Thus, we only report one of the regression result here.

tokens heard					
Step 2				0.218*	0.057
Total number of tokens heard	0	0	0.342		
Total number of times of clearing level 10	0.018	0.013	0.247		

N1=30, N2=30, N3=30, *p<.05

Table 10 shows that the regression models in step 1 and step 2 significantly predicted the d' scores. In step 1, R^2=.161, $F(1, 28) = 5.36$, p<.05. In step 2, R^2=.218, $F(2, 27) = 3.76$, p<.05. Interestingly, adding the total number of times of clearing level 10 as the predictor did not improve the model significantly in step 2. In step 2, ΔR^2=.057, $\Delta F(1, 27)=1.98$, p=.171. In the second model where both predictors were entered, only the total number of input tone tokens was a significant predictor (*SE B*=.386, $t(27)=2.2$, p<.05). Therefore, the first hierarchical regression showed that the total number of input tokens during the video game training significantly predicted the d' scores, namely, the more tokens were heard the better the sensitivity to lexical tones in disyllable context became.

Since the results implied that only the total number of input tone tokens mattered for the sensitivity to lexical tones in disyllables, we ran an additional simple regression for the non-minimal pair disyllable training group to test whether their d' scores can be predicted simply by the total number of input tokens heard by the participants. The result showed that the regression model only with the total number of input tokens as the predictor just reached the significance level: R^2=.387, $F(1,8)=5.1$, p=.05. The near significance result was likely due to the small sample size because there were only ten

participants' data points in the simple regression model. The reason we did not combine the non-minimal pair disyllable training group's result with the other three training groups' results is that the word identification accuracy rate (game accuracy) and the total number of times of clearing level 10 reached ceiling for the non-minimal pair training group. The ceiling effect reached by the non-minimal pair disyllable training group was due to the fact they were simply using the first syllables or the segmental information to play the game. Nevertheless, the result of the simple regression for the non-minimal pair disyllable training group supported the claim that the total number of input tokens can predict the sensitivity to lexical tones in the disyllable context.

4.5.2 Relation between game performance and word identification for old talker stimuli

The second hierarchical regression used the same two predictors— (1) the total number of tone tokens heard by the participants during the four days' video game playing; (2) the total number of times of clearing level 10 to predict the word/tone identification accuracy rates for old talker stimuli. Since the word identification accuracy rate result is dichotomous, we transformed the accuracy rate into logit, using stepwise logistic regression. We entered the two predictors in steps. The result is summarized in Table 11.

Table 11. Hierarchical logistic regression result with two predictors: total number of tone tokens heard during the video game training and total number of times of clearing level 10. DV: word/tone identification accuracy rates transformed into logit for old talker stimuli.

	SE B	β (standardized B)	AIC	χ^2
Dependent variable:	Word accuracy rate transformed into logit			
Step 1			186.92	
Total number of tokens heard	0.1347	1.0129		
Step 2			179.16	9.76*
Total number of tokens heard	0.142	1.005		
Total number of times of clearing level 10	0.016	0.05		

N1=30, N2=30, N3=30, df=1 for χ^2, *p<.05, ** p<.01, *** p<.001

As Table 11 shows, adding the total number of times of clearing level 10 significantly improved the model (χ^2=9.76, *df*=1, *p*<.05). In the model with two predictors, the total number of input tokens was a significant predictor (β=-2.48E-04, *z*= -1.974, *p*<.05) and total number of times of clearing level 10 rate was a significant predictor (β=1.08, *z*=6.892, *p*<.001). The results suggested that the more tokens the participants heard, the higher the probability of correct response for the old talker stimuli would be; the more times of clearing level 10, the higher the probability of correct response for the old talker stimuli would be.

4.5.3 Relation between game performance and word identification for new talker stimuli

The third hierarchical regression used the same two predictors—(1) the total number of tone tokens played during the four days' video game playing; (2) the total

number of times of clearing level 10 to predict the word/tone identification accuracy rates transformed into logit for new talker stimuli. Again, we ran a stepwise logistic regression. The result is summarized in the Table 12.

Table 12. Hierarchical logistic regression result with two predictors: total number of tone tokens heard during the video game training and total number of times of clearing level 10. DV: word/tone identification accuracy rates transformed into logit for new talker stimuli.

	SE B	β (standardized B)	AIC	χ^2
Dependent variable:	Word accuracy rate transformed into logit			
Step 1			276.47	
Total number of tokens heard	5.31E-05	2.04E-04		
Step 2			265.94	12.5**
Total number of tokens heard	5.47E-05	1.604E-04		
Total number of times of clearing level 10	1.255E-02	4.4E-02		

N1=30, N2=30, N3=30,df=1 for χ^2, *p<.05, ** p<.01, *** p<.001

Table 12 shows that adding the total number of times of clearing level 10 significantly improved the model (χ^2=12.5, *df*=1, *p*<.05). In the model with two predictors, only the total number of times of clearing level 10 was a significant predictor (β=0.735, *z*=5.78, *p*<.001). The result suggested that the total number of times of clearing level 10 was the only predictor for the ultimate word identification accuracy rate for the new talker stimuli.

Since the total number of input tokens and the total number of times of clearing level 10 were found to be the significant predictors for the word identification accuracy rate for the old talker stimuli and the new talker stimuli, we added these predictors to the previous mixed effect logistic regression model described in Section 4.4 to examine whether the model was significantly improved. The result showed an improvement with marginal significance (χ^2=6.3, *df*=2, *p*=0.06). Thus, it suggested that the majority of the variance of log odds can be accounted for by factors Talker, Tone and Training. The total number of input tokens and total number of times of clearing level 10 may be secondary predictors for the word identification accuracy rate.

4.5.4 Summary of the relation between game performance and tone categorization performance

Four trainee groups received a large number of tone tokens as input during the four days of video game training. Each participant heard at least 4000 tone tokens with an approximately equal number of tokens for each lexical tone during the training. The three trainee groups trained with words with contrastive tones, namely, words that only differed in terms of lexical tones showed different rates of reaching the ceiling. The individuals also differed in terms of the amount of input and the total number of times of clearing the final level of the game. Hierarchical Regression results suggested that only the total number of input tone tokens can predict the sensitivity to lexical tones in the disyllable context. In terms of predicting the word/tone identification accuracy rate for the old talker stimuli, the total number of tokens and the total number of times of clearing

the final level were the two significant predictors. In terms of predicting the word/tone identification accuracy rate for the new talker stimuli, only the total number of times of clearing the final level was a significant predictor.

4.6 Relation between cue-weighting and tone categorization performance

We also examined the relation between the cue-weighting on the pitch direction dimension and the d' scores of tone discrimination in disyllables and the word identification accuracy rate in the posttest. Neither the absolute cue-weighting values on pitch direction dimension nor the difference between the posttest weighting values and the pretest weighting values was a significant predictor for any of the tone categorization parameters. In spite of the absence of a direct quantitative relation between cue-weighting and tone categorization performance, we found a qualitative relation between cue-weighting and tone discrimination, as shown in Section 4.3.2, namely, only the training groups that had a cue-weighting increase on pitch direction showed a d' increase for T2-T4 in monosyllables in the posttest; groups that did not have a cue-weighting increase on pitch direction did not show a d' increase for T2-T4 in monosyllables.

The null results of the regression analysis that examined the quantitative relation between cue-weighting and tone categorization performance only excluded the potential linear relationship between cue-weighting and tone categorization. It did not necessarily mean cue-weighting shift towards pitch direction dimension had no effect for tone categorization. It seems that the relation may be quite complicated.

First of all, the variance-manipulated training group and the multi-talker training group had a cue-weighting increase on pitch direction after the training whereas the minimal pair disyllable training group and the non-minimal pair disyllable training group did not, as shown in Section 4.2. Despite the absence of cue-weighting increase on pitch direction, the minimal pair disyllable training group and the non-minimal pair disyllable training group both had a d' score increase for tone discrimination in the disyllable context. These results suggest that the cue-weighting increase on pitch direction may not directly lead to an increase in the sensitivity to lexical tones in disyllable context. As long as there was a sufficient number of input tone tokens, it would help improve tone discrimination in the disyllable context. It is worth noting that, even for the non-minimal pair disyllable training group, there was a significant d' increase in the disyllable context. Even though the participants in the non-minimal pair disyllable training group only paid attention to the segmental information instead of the lexical tones on the second syllable, their sensitivity to lexical tones in disyllables still increased. These results suggest that the amount of lexical tone input might be more directly related to discriminative sensitivity improvement than cue-weighting shift towards pitch direction.

Second, the multi-talker training condition turned out to be the most robust for shifting cue-weighting towards pitch direction relative to the other three training conditions as the multi-talker training group had a larger cue-weighting increase on pitch direction dimension than the other training conditions. Moreover, among all four trainee groups, only the multi-talker training group had a significant cue-weighting decrease on pitch height dimension. The multi-talker training group outperformed the other three

training groups with a larger d' increase for tone discrimination in the disyllable context and a better generalization of word/tone identification to new talker stimuli. It could be the case that only with a cue-weighting increase on pitch direction and a cue-weighting decrease on pitch height can a higher sensitivity to lexical tones be generated and a better word/tone identification be generalized to new talkers. However, we cannot simply attribute the multi-talker training group's better generalization for word/tone identification to new talker stimuli to the increased cue-weighting on pitch direction and the decreased cue-weighting on pitch height as the multi-talker training condition included more indexical information (e.g., gender, age, voice quality, etc.) than other three training groups. All these indexical information may contribute to the tone categorization improvement as well. More on the potential usefulness of indexical information for sound categorization is discussed in Chapter Five.

CHAPTER FIVE: DISCUSSION

In this chapter, we discuss the three themes that the current study set out to investigate. In Section 5.1, we compare the training efficiency of our video-game training paradigm to a few training paradigms used in previous tone training studies. In Section 5.2, we discuss the effect of talker variability, variance, minimal pairs and non-minimal pairs for shifting cue-weighting for tone perception. In Section 5.3, we discuss the effectiveness of different types of training stimuli on L2 tone categorization (e.g., tone identification of both old and new talkers; tone discrimination). In Section 5.4, we discuss some implications of the current study for theories of sound categorization.

5.1 Video-game training efficiency

One of the goals of this study is to investigate whether a multi-modal phonetic training paradigm like the video-game training paradigm can help naïve listeners learn new sound categories more efficiently. Before delving into the efficiency issue, we want to know whether the learners indeed formed four lexical tone categories after the training. One criterion for evaluating the tone category learning is the word/tone identification accuracy rate. The following two tables show the accuracy rates of the three groups who were trained on words with contrastive tones.

Table 13. Word/Tone identification accuracy rate (%) of the variance-manipulated training, multi-talker training and minimal pair disyllable training groups for the old talker stimuli.

old talker	T1	T2	T3	T4	overall
Multi-talker	66.25	71.25	90	70	74.38
Variance-manipulated	78.75	81.25	68.75	66.25	73.75
Disyllable minimal pair	96.25	91.25	96.25	68.75	88.13

Table 14. Word/Tone identification accuracy rate (%) of the variance-manipulated training, multi-talker training and minimal pair disyllable training groups for the new talker stimuli.

new talker	T1	T2	T3	T4	overall
Multi-talker	77.50	75.00	82.50	80.00	78.75
Variance-manipulated	81.25	66.25	50.00	52.50	62.50
Disyllable minimal pair	57.50	62.50	56.25	52.50	57.19

As Table 11 shows, after the video-game training, the three training groups' overall tone identification accuracy rate was well above chance level for the old talker stimuli. In terms of the accuracy rate for different lexical tones, all three training groups identified the four lexical tones consistently well above chance level, especially the disyllable minimal pair training group. For the new talker stimuli, as shown in Table 12, the variance-manipulated training group and the disyllable minimal pair training group had lower overall accuracy rates than the multi-talker training group but the word identification accuracy rates of the three groups were still above chance level. The reason that the disyllable minimal pair training group not only had the lowest overall word

identification accuracy rate but also low accuracy rates across the four lexical tones is likely due to the tonal coarticulation in the test stimuli particularly for this training group. The disyllable minimal pair training group was trained with disyllables that concatenated T1 and four resynthesized lexical tones used in the variance manipulated training. Even though we smoothed out the pitch transition between the preceding T1 and the following tones by adjusting the pitch offset of the preceding T1, the pitch transitions in the minimal pair disyllables in the training stimuli still may not be the same as the naturally produced pitch transitions between two tones in the test stimuli. It is possible that the lack of natural tonal coarticulation in the training stimuli caused the lower word identification accuracy rates for new talker stimuli.

In the perceptual training of Mandarin Chinese tones of Wang et al (1999), the participants were all native English speakers who had already learned Mandarin Chinese for at least 4 months in a classroom setting. Their average tone identification accuracy rate before the multi-talker training was 69%. Our implicit tone category learning paradigm generated a comparable tone identification accuracy rate to that achieved through almost half a year's formal Mandarin Chinese learning. However, in Wang et al. (1999), the syllable types used for tone identification were highly variable whereas in our current study, we were only using either the monosyllable *yu* or the disyllable *ta1yu* for the identification task. Thus, our higher accuracy rate than the one in Wang et al (1999) was likely due to the simpler syllable structure we used for the tone identification task. Nevertheless, four different tone categories were indeed established by the three training groups after the implicit word learning paradigm. The total amount of time that took the

participants to learn the four tone categories was only 1 hour and 45 minutes (30 mins for Day 1 to 3. On Day 4, the participants only played the game 15 mins before the posttest) in four days. In addition, within such a short time period, the participants' sensitivity to lexical tones in both monosyllable and disyllable contexts also significantly improved. The naïve listeners' tonal discrimination performance in monosyllable context in Chandrasekaran et al (2010) had an average d' of 4.1. Our training result was comparable to their result as two of our training groups' average d' reached near 4.0 and two training groups' d' were over 4.0. However, their training period lasted 9 sessions, each of which lasted 30 minutes with a total of four and half hours. Since our study synthesized the four Mandarin tones for the speeded AX discrimination task with the same parameters used in Chandrasekaran et al. (2010), it suggests that the tone discrimination result in the monosyllable context achieved by the naïve listeners in the current study is comparable to the result in Chandrasekaran et al. (2010), but the result was achieved with much less training time.

The generally comparable tone categorization results in the current study to the previous tone training studies may come from two sources: the multi-modal phonetic training paradigm and the number of training stimuli. The first source may be related to the learning motivation caused by the intrinsic reward during the video-game play. Although there was no explicit feedback provided to the participants during the video-game training, we did use visual information to let the participants know whether their choices of food to feed the animal is correct or incorrect. For example, an animal disappears when it gets its favorite food and an animal keeps moving on the screen when

it does not get its favorite food. Therefore, the implicit learning here is a type of semi-supervised learning, namely, participants receive feedback indirectly. Studies in neuroscience found that when a learning process involves a paradigm where intrinsic rewards are provided (e.g., clearing difficult levels in a video game if participants are able to identify lexical tones in the current study), the striatal reward system is activated during the learning process and may further motivate the learning. With the learning motivation, participants will be able to learn the lexical tones more efficiently. In terms of the number of training stimuli, Wang et al. (1999) had participants attend eight training sessions. It took two sessions to expose a participant to 180 tone stimuli produced by one talker, and there were six talkers (3 males and 3 females). Thus, in total, each participant was trained with 4320 tone tokens (180 tone stimuli x 6 talkers x 4 consecutive sessions). Chandrasekaran et al. (2010) had participants attend nine training sessions. In each training session, there were 24 words contrasted by lexical tones, each of which was produced by four talkers and repeated four times. Thus, in total, there were 3456 training tone tokens (24 words x 4 talkers x 4 repetitions x 9 sessions). In the current study, each participant was exposed to over 4000 tone tokens. Therefore, the amount of training tone tokens used in the current study is similar to the ones used in the previous studies. Because of the comparable number of training stimuli used in the current study and the previous studies, we cannot make a strong claim about the learning efficiency that could potentially be elicited by the video-game training paradigm. But since there was no explicit feedback in the video-game training, it saved a significant amount of time for the participants to be exposed to the training stimuli.

The tone categorization performance comparisons made between our study and other studies may not be very precise as the training and test conditions were not exactly the same. Nevertheless, within 1 hour and 45 minutes, our training conditions allowed naïve listeners to identify different lexical tones and increase their sensitivity to lexical tones in both monosyllable and disyllable contexts.

5.2 Effect of talker variability and variance manipulation on cue-weighting for tone perception and its relation to tone categorization

The second goal of the current study is to examine the effect of talker variability and the effect of manipulating the variance on different acoustic dimensions on the perceptual cue-weightings. In the current study, we essentially used two types of training stimuli. The first type was the resynthesized tone tokens that did not overlap on the pitch direction dimension but had a larger overlap on the pitch height dimension. In terms of jnd, there was a larger variance on pitch height than on pitch direction. The second type of training stimuli was the multi-talker tone tokens that had overlap both on the pitch direction and pitch height dimensions but with a much larger overlap on pitch height. Consistent with Chandrasekaran et al.'s (2010) finding that multi-talker tone tokens (two males and two females) in an implicit tone training paradigm made naïve listeners (native English speakers in their case) shift more weight towards the pitch direction dimension, the primary acoustic dimension native Chinese speakers rely on for tone perception, our study also showed that multi-talker tone tokens helped naïve listeners (native English speakers) shift more weight towards pitch direction after the video game training, which

was also an implicit tone learning paradigm. Importantly, our manipulation of the variance on the pitch direction and pitch height dimensions also helped naïve listeners shift their cue-weighting towards the pitch direction dimension. These results are largely consistent with the results found in Holt and Lotto (2006) and Lim and Holt (2011) that the cue-weighting is shifted towards the acoustic dimension that has a smaller variance. However, at this point we cannot make a very strong claim about the effectiveness of variance-manipulation in terms of shifting cue-weighting at the suprasegmental level. The reason is that although the variance on pitch direction was smaller than that on pitch height in terms of jnd, it is still possible that the theoretical just noticeable difference between two tone tokens within the same lexical tone category cannot be heard by the naïve listeners. Because the jnd for discriminating the synthesized pitch contours found in the psychophysics studies may not fully apply to the discrimination for naturally produced pitch contours. Thus, we need to be cautious about claiming that the smaller variance on pitch direction made naïve listeners shift their cue-weighting towards pitch direction.

Another point worth mentioning is that the overlap on the pitch direction dimension among the training tokens in the multi-talker training seemed not to hamper the cue-weighting shift towards pitch direction. Although there was no overlap on the pitch direction dimension among the training tokens in the variance-manipulated training, its training effect was not as robust as the multi-talker training in terms of shifting cue-weighting towards pitch direction. Thus, it seemed that the overlap on pitch direction among the training tokens did not hamper the cue-weighting shift. This suggests that

sound categorization in a multi-dimensional acoustic space may not need to have a dimension in which the categories in question are completely distinct from each other. An optimal sound classification should allow some degree of overlap between sound categories in the acoustic space.

The result of the discrimination for specific tone pairs in the monosyllable context suggests a relation between tone discrimination and cue-weighting, namely, only the variance-manipulated training and multi-talker training groups that had a cue-weighting increase on pitch direction showed a d' increase for discriminating T2 and T4 after the training. Other training groups, including the non-native control group, that did not have a cue-weighting increase on pitch direction did not show a d' increase for discriminating T2 and T4. These results suggest that the discrimination for specific tone pairs is related to the cue-weightings on pitch height and pitch direction as only the participants in the variance-manipulated training and multi-talker training conditions shifted their cue-weighting towards pitch direction and their perceptual distance between T2 and T4 increased, thus, their discrimination for T2-T4 improved as well.

Another finding in terms of cue-weighting is that individual preferences on using pitch direction and pitch height varied within both native Chinese speakers and native English speakers. Similar results were also found in Gandour (1983). But as a language group, native Chinese speakers and native English speakers differed in terms of which acoustic cue is the primary dimension for tone perception. Due to the existence of individual differences, the participants in the variance-manipulated training group had a smaller variability than the participants in the multi-talker training group in terms of cue-

weighting on pitch direction and pitch height dimension before the training. Despite the smaller degree of homogeneity of cue-weighting among the individuals in the multi-talker training group relative to the individuals in the variance manipulated training group, the multi-talker training group had a larger cue-weighting increase on pitch direction than the variance manipulated training group after training. Therefore, multi-talker training seemed to be more robust for the cue-weighting shift towards pitch direction than variance-manipulated training in terms of overcoming the larger individual cue-weighting variability.

5.3 The effect of sound input distribution on sound discrimination

In terms of the tone discrimination results, a crucial finding was that regardless of training condition, the sensitivity to lexical tones in both monosyllable and disyllable contexts significantly improved for all four trainee groups. The non-native control group also had a d' increase for tone discrimination in the monosyllable context, indicating a practice effect for such a task. But for tone discrimination in the disyllable context, the control group did not have any d' increase, suggesting that tone discrimination in the disyllable context is a more demanding task. The result of the consistent d' increase among the training groups was consistent with the distributional learning theory that the input was clustered into distinct categories based on frequency tracking of the stimuli input. Such clustering helped sound discrimination. For example, Maye et al. (2000) and Feldman et al. (2011) both found sound discrimination improvement among the training

groups as long as the distinct sound categories were equally distributed in the sound input. In addition, our study showed that the more input there was, the better the tone discrimination was, suggesting that the amount of input was a crucial factor for tone discrimination improvement.

Another goal of the current study was to examine whether non-minimal pair training can generate better sensitivity to lexical tones than minimal pair training in both monosyllable and disyllable contexts. Our pilot study found that the naïve listeners in the non-minimal pair training condition had a sensitivity increase for T2-T3 discrimination whereas the participants in the minimal pair training condition did not have any sensitivity increase, but the current study showed that the participants in the non-minimal pair training condition had a comparable amount of d' increase as the minimal pair training condition in both monosyllable and disyllable contexts. Thus, overall, the non-minimal pair training condition was not better than the minimal pair training condition. However, there is a crucial difference between the current study and the pilot study. In the pilot study, the participants did not need to do anything during the familiarization phase. All they needed to do was to listen to the training stimuli. After two familiarization phases, the non-minimal pair training group turned out to have a larger sensitivity increase for two acoustically close tone categories than the minimal pair training group. The pilot study's result showed that, without any form of feedback, only the non-minimal pair training group had tone discrimination improvement whereas the minimal pair training group did not have any tone discrimination improvement. To account for these results, we argued that the non-minimal pair training condition (e.g.,

ku1ju2 vs. *po1ju3*) biased the participants towards two tone categories and the minimal pair training condition (e.g., *ku1ju2* vs. *ku1ju3*) biased the participants towards one tone category. In the end, the bias for two tone categories caused improvement for tone discrimination between *ju2* and *ju3* for the non-minimal pair training condition. In the current study, even though there was no explicit feedback that told the participants which tone category corresponded to which animal, the participants still received positive feedback when the correct food was selected to feed the animal and the animal disappeared. The participants also received negative feedback when the wrong food was selected to feed the animal and the animal kept flashing on the screen. Because of the implicit feedback, in practice, the participants in the non-minimal training condition reported that they only used the first syllable to play the video game and completely ignored the second syllables that carried the contrastive tones. We argue that the unbalanced attention for the segment information and the tone information in the non-minimal pair disyllables made the non-minimal pair training condition lose its advantage in improving tone discrimination relative to the minimal pair training condition. However, despite the lack of attention to the syllables with contrastive tones, the non-minimal pair training group's tone discrimination still significantly improved in both monosyllable and disyllable contexts. This result seems to support the distributional learning theory that the implicit frequency tracking of the four tone clusters with uniform probability in the input led to better discrimination among the four tones.

5.4 Theoretical implications

One interesting finding in the current study is that despite the cue-weighting shift towards pitch direction, the variance-manipulated training group trained with the resynthesized tokens seemed not to be able to generalize the tone identification to naturally produced stimuli whereas the multi-talker training group seemed to be able to make tone identification generalization (variance manipulated training group's accuracy rate for the new talker stimuli: 64%; multi-talker training group's accuracy rate for the new talker stimuli: 79%). The result of the relatively poor word identification generalization to natural stimuli for participants who were trained on the resynthesized stimuli is not uncommon. Lim and Holt (2011) found that the native Japanese speakers who were trained on the resynthesized /r/ and /l/ only exhibited a trend in improving the recognition of naturally spoken /r/-/l/ words in L2 English but did not reach significance level. They argued that the lack of significant improvement from the training with resynthesized tokens may be due to the fact that listeners' performance was already above chance (50%) in pretest as well as individual differences in performance. They argued that the trend of improvement for the identification of natural stimuli suggested that the learning with stylized synthetic speech may have implications for natural spoken word recognition. Since in the current study we only had a posttest word/tone identification, we cannot examine whether the tone identification for natural stimuli improved. But overall it seems that the resynthesized training stimuli have limitations for sound categorization in generalizing to new talker stimuli.

Comparing the tone identification result between the variance-manipulated training group and the multi-talker training group, we can see that the multi-talker training was more robust in terms of generalizing word identification to new talkers. We argue that the mere cue-weighting shift towards pitch direction may not be enough for good tone identification generalization to natural stimuli. Any naturally produced sound is a multidimensional acoustic signal. In terms of syllables with contrastive tones, the pitch height and pitch direction dimensions are only two of the multiple acoustic dimensions of the syllables. Good tone identification may require listeners to use the relevant or reliable acoustic cues in the presence of multiple irrelevant or less reliable acoustic cues. Rost and McMurray (2009) demonstrated the importance of irrelevant acoustic cues for infants to learn words that contrast only in terms of VOT. They showed that 14 months old infants who were exposed to exemplars of the minimal pair (/buk/ and /puk/) produced by multiple speakers successfully associated the sounds with the visual objects whereas the infants who were exposed to exemplars of the minimal pair produced by a single speaker did not show any sound-to-object association. To account for the robustness of the multi-talker training for learning minimal pairs, Rost and McMurray argued that there were at least two kinds of relevant variability and hence two kinds of learning mechanisms that may be important for learning minimal pairs. One is the variability along specifically phonetic dimensions (e.g., pitch height, pitch direction). The other is the variability in non-phonetic information, which may help learners extract the relatively invariant phonetic dimensions. The first type of variability may allow the learners to define the phonetic or lexical categories that contrast the words. This would

require distributional learning mechanisms (Maye et al. 2002; see also Maye et al., 2008). This approach posits that learners track the frequencies of specific phonetic cues and extract categories from the natural clusters. The second type of variability, which is the variability in irrelevant aspects of the stimuli may improve the learning of contrastive sounds by paying attention to those aspects of the input that are comparatively stable. As Rost and McMurray (2009) showed in their study, measurements of pitch and the first four formants (measurements of vowel quality) of multi-talker stimuli were all highly variable. Most importantly, none of those cues differed significantly between /buk/ and /puk/, suggesting that they would not be available to directly contrast the words. Nonetheless, the immense amount of irrelevant variation present would provide the necessary redundancy for the sort of learning mechanism that uses non-critical variation to extract the invariant elements from a noisy signal. Back to the cue-weighting results and tone identification result found in the current study, the robustness of the multi-talker training relative to other training conditions using resynthesized tokens was likely due to the fact that the multi-talker training included both relevant phonetic information and irrelevant non-phonetic information. Thus, multi-talker training allowed participants to shift their cue-weighting more towards pitch direction than variance-manipulated training and the multi-talker training had the best tone identification generalization to new talkers.

Relating the usefulness of non-phonetic information to the cue-weighting shift results that occurred in the two monosyllable training groups (the variance manipulated training group and the multi-talker training group) in the current study, only the multi-talker training group had a significant cue-weighting decrease on pitch height. Therefore,

it seemed that the multi-talker training group was reorganizing its perceptual space whereas the variance manipulated training group was simply expanding its perceptual space. The training stimuli used for the variance manipulated training group and the multi-talker training group both had a significant overlap on the pitch height dimension. At first, we expected the resynthesized tone tokens used in the variance manipulated training to lead to cue-weighting decrease on pitch height. But the participants in the variance manipulated training still honed in on the pitch height after the training whereas the participants in the multi-talker training started reducing their dependence on pitch height. Thus, it seemed that the multi-talker training was more efficient in terms of making the learners realize that the pitch height is a secondary acoustic cue for tone perception. We argue that the large variance on pitch height caused by speaker variability better facilitated learners' identification of the importance of the pitch direction than the variance created in the resynthesized stimuli.

Another important finding in the current study is that the relatively poor tone discrimination performance in the disyllable context. There was tonal coarticulation in the disyllable test stimuli. However, such natural tonal coarticulation was missing in the training. Even though for the disyllable training conditions, we concatenated a high level tone with the variance-manipulated lexical tones by shifting the offset of the preceding tone to be closer to the onset of the following tone in order to mimic the natural tonal coarticulation, the resynthesized disyllables may still lack the acoustic characteristics of the naturally produced tonal coarticulation. Thus, the difference between the syllable structures used in the training and the test for the monosyllable training groups and the

lack of natural tonal coarticulation in the training for the disyllable training groups may explain the trainees' poorer tone categorization in the disyllable context. This result suggests that the acoustic information that comes from the contextual variability is important for learning sound categories. Several studies have already shown the effect of contextual variability on tone categorization For example, Moore and Jongman (1997) showed that the average f0 and the pitch offset of the preceding tone biased native Chinese speakers' tone identification for the synthesized tones, which were gender ambiguous. Sereno et al. (2012) systematically studied the effect of speaking rate on tone identification and showed different speaking rates of the precursor sentences biased native Chinese speakers' identification of a tone continuum from T2 to T3. In terms of computational modeling of the effect of contextual variability on sound categorization, McMurray and Jongman (2011) studied the informational assumptions of several models of speech categorization, in particular, the number of cues that are the basis of categorization and whether these cues represent the input veridically or have undergone compensation. A corpus of 2880 American English fricative productions (Jongman, Wayland & Wong, 2000) spanning many talker- and vowel-contexts was used and 24 cues for each fricative were measured. A subset was also presented to listeners in an 8AFC phoneme categorization task. The researchers trained a common classification model based on logistic regression to categorize the fricative from the cue values, and manipulated the information in the training set to contrast 1) models based on a small number of invariant cues; 2) models using all cues without compensation, and 3) models in which cues underwent compensation for contextual factors. Compensation was

modeled by Computing Cues Relative to Expectations (C-CuRE), a new approach to compensation that preserves fine-grained detail in the signal. Only the model with compensation (e.g., gender: expected f0 of male and female) achieved a similar accuracy to listeners, and showed the same effects of context. The researchers argued that sound categorization can overcome the variability in speech when sufficient contextual information is available and some form of compensation schemes is employed.

Relating the tone categorization found in the current study to the research that showed the importance of non-phonetic cues and contextual variability for sound categorization, we argue that the indexical and contextual information play crucial roles for tone categorization as well. The importance of these types of information echoes the Exemplar model for sound categorization, which claims that the detailed acoustic information of each individual input signal either in isolation or in fluent speech is stored in memory for sound categorization.

CHAPTER SIX: CONCLUSION

The current study implemented a multi-modal phonetic training paradigm, namely, video game training, for naive listeners' learning of Chinese tone categories. The result showed that the video game training paradigm was highly efficient for naïve listeners (native English speakers in this case) to form four tone categories. Within less than two hours of video game training, all trainees reach a tone identification accuracy rate well above the chance level.

Different training conditions have different effects on the participants' cue-weighting. The two disyllable training groups did not show any cue-weighting change on the pitch direction and pitch height dimensions while the two monosyllable training groups showed a cue-weighting shift towards pitch direction, the acoustic dimension native Chinese speakers primarily rely on for tone perception. The multi-talker training group showed more cue-weighting shift towards the pitch direction dimension than the variance manipulated training group. In addition, only the multi-talker training group showed a cue-weighting decrease on pitch height, which is a less reliable cue for tone categorization. Based on these results, we argue that the manipulation of variance on pitch direction and pitch height dimension is able to shift cue-weighting towards the pitch direction dimension, but the multi-talker training condition is more robust than the variance manipulated training condition in terms of adjusting the cue-weighting to be more nativelike.

Talker variability is not only effective in shifting native English speakers' cue-weighting for tone perception to be more nativelike, but also effective in improving tone categorization, specifically, generalizing tone identification to new talkers. However, multi-talker training may not be able to allow naïve listeners to generalize tone discrimination from the monosyllable to the disyllable context. These results suggest the importance of indexical and contextual information for sound categorization.

Finally, the tone discrimination results support the distributional learning theory that implicit frequency tracking of distinct sound categories leads to better sound discrimination. First, we found that the total amount of tone input was a significant factor in predicting the ultimate tone discrimination performance, namely, the more input there is, the better the tone discrimination becomes. Second, despite the lack of attention to the contrastive tones in the non-minimal pair disyllable training condition, the tone discrimination still significantly improved, suggesting that implicit word learning occurred during the implicit statistical learning of the four tone categories.

REFERENCES

Abramson, A. S. (1962).The vowels and tones of standard Thai: Acoustical measurements and experiments (Vol. 20). Bloomington, IN: Indiana University Research Center in Anthropology, Folklore, and Linguistics.

Arabie, Phipps. Carroll, J. Douglas. and DeSarbo, Wayne.(1987) Three-way scaling and clustering Newbury Park ; London : SAGE.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. Psychological Review, 95, 124 –150.

Best, C. T., McRoberts, W., and Goodell, E (1988) Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. J Acoust Soc Am. 2001; 109(2): 775–794.

de Boer, Bart (2000) Self organization in vowel systems, Journal of Phonetics 28 (4), 441–465.

Borg, Ingwer. Groenen, Patrick J. F. (2005) Modern multidimensional scaling theory and applications (2nd ed). New York: Springer.

Bortfeld, Heather, Morgan, James L., Golinkoff, Roberta M., & Rathbun, Karen (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. Psychological Science, 16 , 298-304.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., and Tohkura, Y.(1997).Training the Japanese listener to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production, J. Acoust. Soc. Am. 101, 2299–2310.

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., & Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language contrast. NeuroImage, 19, 113–124.

Carroll, J. D., and Chang, J. J. (1970). "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," Psychometrika35, 283–319.

Chandrasekaran, B., Gandour, J. T., and Krishnan, A. (2007a). "Neuroplasticity in the processing of pitch dimensions: A multidimensional scaling analysis of the mismatch negativity," Restor. Neurol. Neurosci. 25, 195–210.

Chandrasekaran, B., Krishnan, A., and Gandour, J. T. (2007b). "Mismatch negativity to pitch contours is influenced by language experience," Brain Res. 1128, 148–156.

Chandrasekaran, B., Sampath, P. D., and Wong, P. C. (2010). Individual variability in cue-weighting and lexical tone learning. J Acoust Soc Am.128 (1), 456-465.

Coster, D. C., & Kratochvil, P. (1984). Tone and stress discrimination in normal Peking dialect speech. In B. Hong (Ed.), New papers in Chinese linguistics(pp. 119–132). Canberra: Australian National University Press.

Delgado, M. R., Stenger, V. A., & Fiez, J. A. (2004). Motivation-dependent responses in the human caudate nucleus. Cerebral Cortex, 14(9), 1022–1030.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). Learning phonetic categories by learning a lexicon. Proceedings of the 31st Annual Conference of the Cognitive Science Society.

Feldman, N. H., Myers, E., White, K., Griffiths, T. L., &Morgan, J. L. (2011). Learners use word-level statistics in phonetic category acquisition. Proceedings of the 35th Boston University Conference on Language Development.

Francis, A. L., Ciocca, V., Ma, L., and Fenn, K. (2008). "Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers," J. Phonetics 36, 268–294.

Gandour, J. (1983). "Tone perception in Far Eastern languages," J. Phonetics 11, 149–175.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. Hearing Research, 47, 103–138.

Goudbeek, M., Smits, R., Cutler, A., & Swingley, D. (2005). Acquiring auditory and phonetic categories. In H. Cohen & C. Lefebvre (Eds.), Categorization in cognitive sciences (pp. 497–513). Amsterdam: Elsevier.

Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. Speech Communication, 50, 109–125.

Green, C. S., & Bavelier, D. (2007). Action videogame experience alters the spatial resolution of attention. Psychological Science, 18(1), 88–94.

Hallé, P. A., Chang, Y.-C, & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. J. Phonetics, 32 , 395–421.

Holt, L. L. & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. J Acoust Soc Am., 119, 3059-3071.

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

Huang. Ts. (2001) Tone perception by speakers of Mandarin Chinese and American English. The Interplay of Speech Perception and Phonology, OSUWPL, vol. 55.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. Cognition, 87, B47–B57.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching /r/ -/l/ to Japanese adults. J Acoust Soc Am., 118, 3267 –3278.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives, J Acoust Soc Am., 106, 1252-1263.

Jusczyk, P. (1993a). "Infant speech perception and the development of the mental lexicon," in The Transition from Speech Sounds to Spoken Words: The Development of Speech Perception, edited by H. C. Nusbaum and J. C. Goodman (MIT, Cambridge, MA).

Jusczyk, Peter W., &Aslin, Richard N. (1995). Infants' detection of the sound patterns of words in fluent speech. Cognitive Psychology, 29, 1-23.

Kaan, E., Wayland, R., Bao, M., and Barkley, C. M. (2007). "Effects of native language and training on lexical tone perception: An event-related potential study," Brain Res. 1148, 113–122.

Keating, P.A., Esposito, C. 2006. Linguistic voice quality. Paper presented at the 11th Australasian International Conference on Speech Science and Technology.

Klein, D., Zatorre, R. J., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. NeuroImage, 13, 646–653.

Koepp, M. J., Gunn, R. N., Lawrence, A. D., Cunningham, V. J., Dagher, A., Jones, T., Brooks, D. J., Bench, C. J., & Grasby, P. M. (1998). Evidence for striatal dopamine release during a video game. Nature, 393, 266–268.

Kruschke, J. (1992). "ALCOVE: An exemplar-based connectionist model of category learning," Psychol. Rev. 90, 22-44.

Kuhl, P. K. (1991a). "Human adults and human infants show a 'perceptual magnet effect' for the prototype of speech categories, monkeys do not," Percept. Psychophys. 50, 93-107.

Kuhl, P. K. (1991b). "Speech prototypes: Studies on the nature, function, ontogeny and phylogeny of the 'centers' of speech categories," in Speech Perception, Production and Linguistic Structure, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (OHM, Tokyo), pp. 239-264.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). "Linguistic experience alters phonetic perception in infants by 6 months of age," Science 255, 606-608.

Liljencrants, L. & Lindblom, B. (1972) Numerical simulations of vowel quality systems: the role of perceptual contrast, Language, 48, 839-862.

Lim, S.–J. Lim & Holt, L. L. (2011). Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization. Cognitive Science, 35, 1390-1405.

Lively, S. E., Logan, J. S., and Pisoni, D. B. (1993). “Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environ-ment and talker variability in learning new perceptual categories,” J. Acoust. Soc. Am. 94, 1242–1255.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (1991). “Training Japanese listeners to identify English /r/ and /l/: a first report,” J. Acoust. Soc. Am. 89, 874–886.

Lotto, A. J., Sato, M., and Diehl, R. L. (2004). “Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/,” in J. Slifka, S. Manuel, and M. Matthies (Eds.) From Sound to Sense: 50 Years of Discoveries in Speech Communication (MIT Research Laboratory in Electronics, Cambridge, MA), pp. C-181–C-186.

Macmillan, N. A., and Creelman, C. D. (1991). Detection Theory: A User's Guide. Cambridge University Press, Cambridge.

MacKain, Kristine S. (1982) Assessing the role of experience on infants' speech discrimination. Journal of Child Language, 9, 527-542.

Maye, Jessica, & Gerken, LouAnn (2000). Learning phonemes without minimal pairs. In S. C. Howell & S. A. Fish & T. Keith-Lucas (Eds.), Proceedings of the 24th Annual Boston University Conference on Language Development (pp. 522-533). Somerville, MA: Cascadilla Press.

Maye, Jessica, Werker, Janet F., & Gerken, LouAnn (2002). Infant sensitivity to distributional information can affect phonetic discrimination. Cognition, 82 , B101-B111.

McMurray, B., and Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review 118, 219-246.

Miller, J. L. (1977). "Properties of feature detectors for VOT: The voiceless channel of analysis," J. Acoust. Soc. Am. 62, 641-648.

Moore, C.B., and Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. J Acoust Soc Am. 102, 1864-1877.

Nosofsky, R. M. (1986). "Attention, similarity, and the identification-categorization relationship," Journal of Experimental Psychology, 115, 39-57.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), Frequency and the emergence of linguistic structure. Amsterdam: John Benjamins.

Pisoni, D. B., Lively, S. E., & Logan, J. S. (1994). Perceptual learning of nonnative speech contrasts: implications for theories of speech perception. In J. C. Goodman & H. C. Nusbaum (Eds.), The development of speech perception: the transition from speech sounds to spoken words (pp. 121 –166). Cambridge, MA: MIT Press.

Repp, B. H. (1976). "Dichotic competition of speech sounds: The role of acoustic stimulus structure," J. Exp. Psychol. Hum. Percept. Perform. 3, 37-50.

Rosch, E. (1973). Natural categories. Cognitive Psychology, 4, 328 –350.

Rost, G., and McMurray, B. (2009) Speaker variability augments phonological processing in early word learning. Developmental Science, 12(2), 339-349.

Sereno, J., H.-J. Lee, & A. Jongman (2011). Perceptual accommodation to rate, talker and context in Mandarin. J Acoust Soc Am. 130(4): 2445.

Seitz, A. R., & Watanabe, T. (2005). A unified model for perceptual learning. Trends in Cognitive Science, 9(7), 329–334.

Seitz, A. R., Kim, D., & Watanabe, T. (2009). Rewards evoke learning of unconsciously processed visual stimuli in adult humans. Neuron, 61, 700–707.

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

Shepard, R. N. (1978). Externalization of mental images and the act of creating. In B. Randhawa & W. Coffman (eds.), Visual learning, thinking, and communication (pp. 133–189). New York: Academic Press.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. Journal of Experimental Psychology, 54, 377– 411.

Strange, W., and Dittmann, S. (1984). "Effects of discrimination training on the perception of/r/-/1/by Japanese adults learning English," Percept. Psychophys. 36, 131-145.

Strange, W. (1995). Cross-language studies of speech perception: A historical review. In W. Strange (Ed.), Speech perception and linguistic experience: Issues in cross- language speech research (pp. 3–45). Timonium, MD: York Press.

Sun, K. C. & Huang, C. (2012) A Cross-linguistic Study of Taiwanese Tone Perception by Taiwanese and English Listeners. JEAL Vol. 21 No. 3, pp.305-327.

Tsushima, Y., Seitz, A. R., & Watanabe, T. (2008). Task-irrelevant learning occurs only when the irrelevant feature is weak. Current Biology, 18(12), R516–R517.

Tricomi, E., Delgado, M. R., McCandliss, B. D., McClelland, J. L., & Fiez, J. A. (2006). Performance feedback drives caudate activation in a phonological learning task. Journal of Cognitive Neuroscience, 18, 1029–1043.

Vallabha, G. K., & McClelland, J. L. (2007). Success and failure of new speech category learning in adulthood: Consequences of learned Hebbian attractors in topographic maps. Cognitive, Affective, & Behavioral Neuroscience, 7, 53–73.

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. J. Acoust. Soc. Am., 118, 2618–2633.

Wang, Y., Spence, M. M., Jongman, A., and Sereno, J. A. (1999). ‘‘Training American listeners to perceive Mandarin tones,’’ J. Acoust. Soc. Am., 106, 3649–3658.

Werker, Janet F., & Tees, Richard C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behavior and Development, 7, 49-63.

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

Wong, P. C., Perrachione, T. K., and Parrish, T. B. (2007). “Neural characteristics of successful and less successful speech and word learning in adults,” Hum. Brain Mapp 28, 995–1006.

Xi, J.; Zhang, L.; Shu, H.; Zhang, Y. and Li, P (2010) Categorical perception of lexical tones in Chinese revealed by mismatch negativity. Neuroscience, Vol. 170 Issue 1, p223-231.

Xu, Y.S., Gandour, J. T., and Francis, A. L. (2006). “Effects of language experience and stimulus complexity on the categorical perception of pitch direction,” J. Acoust. Soc. Am., 120, 1063–1074.

Yamada, R. A. (1995). Age and acquisition of second language speech sounds: perception of American English /r/ and /l/ by native speakers of Japanese. In W. Strange (Ed.),Speech perception and language experience: issues in cross-language research (pp. 305 –320). Baltimore, MD: York Press.

Yip, M. (2002). Tone. Cambridge: Cambridge University Press.

APPENDICES: RESULTS OF MIXED EFFECT LOGISTIC REGRESSION MODELS WITH DIFFERENT BASELINES.

Table A Logistic Regression Analysis of three training groups' word identification accuracy rate, using old talker stimuli, T3 and multi-talker training condition as the baselines.

	β	Std. Error	z value	P (2-tailed)	
(Intercept)	2.3888	0.453	5.273	1.34E-07	***
Talkernew	-0.6748	0.4848	-1.392	0.163975	
Tone1	-1.616	0.4525	-3.571	0.000355	***
Tone2	-1.3621	0.4578	-2.975	0.002926	**
Tone4	-1.4275	0.4562	-3.129	0.001754	**
Trainingvm	-1.5155	0.5724	-2.648	0.008104	**
Trainingminimal	1.1606	0.7944	1.461	0.144024	
Talkernew:Tone1	1.2828	0.6093	2.105	0.03526	*
Talkernew:Tone2	0.8809	0.6091	1.446	0.1481	
Talkernew:Tone4	1.2536	0.6174	2.03	0.042317	*
Talkernew:Trainingvm	-0.1984	0.5943	-0.334	0.738482	
Talkernew:Trainingminimal	-2.5726	0.8119	-3.169	0.001531	**
Tone1:Trainingvm	2.1832	0.5899	3.701	0.000215	***
Tone2:Trainingvm	2.0973	0.6003	3.494	0.000476	***
Tone4:Trainingvm	1.3022	0.5761	2.26	0.023814	*
Tone1:Trainingminimal	1.616	0.9646	1.675	0.093873	.
Tone2:Trainingminimal	0.4309	0.8595	0.501	0.616174	
Tone4:Trainingminimal	-1.2011	0.7986	-1.504	0.132603	

Table B Logistic Regression Analysis of three training groups' word identification accuracy rate, using old talker stimuli, T2 and multi-talker training condition as the baselines.

	β	Std. Error	z value	P (2-tailed)	
(Intercept)	1.0266	0.35378	2.902	0.00371	**
Talkernew	0.20614	0.36866	0.559	0.576051	
Tone1	-0.25382	0.35397	-0.717	0.473345	
Tone3	1.36214	0.4578	2.975	0.002926	**
Tone4	-0.06538	0.35889	-0.182	0.855451	
Trainingvm	0.58192	0.52216	1.114	0.265082	
Trainingminimal	1.5915	0.59475	2.676	0.007452	**
Talkernew:Tone1	0.40184	0.52142	0.771	0.440903	
Talkernew:Tone3	-1.88093	0.60911	-1.446	0.148104	*
Talkernew:Tone4	0.37268	0.53093	0.702	0.482716	
Talkernew:Trainingvm	-1.0667	0.53335	-2	0.0455	*
Talkernew:Trainingminimal	-2.22153	0.60364	-3.68	0.000233	***
Tone1:Trainingvm	0.08583	0.54113	0.159	0.873972	
Tone3:Trainingvm	-2.09735	0.60031	-3.494	0.000476	***
Tone4:Trainingvm	-0.7952	0.52664	-1.51	0.131057	
Tone1:Trainingminimal	1.18514	0.80904	1.465	0.142955	
Tone3:Trainingminimal	-0.43081	0.85955	-0.501	0.616225	
Tone4:Trainingminimal	-1.63191	0.60121	-2.714	0.00664	**

Table C Logistic Regression Analysis of three training groups' word identification accuracy rate, using old talker stimuli, T1 and multi-talker training condition as the baselines.

	β	Std. Error	z value	P (2-tailed)	
(Intercept)	0.77279	0.34629	2.232	0.025638	*
Talkernew	0.60798	0.36884	1.648	0.099274	
Tone2	0.25382	0.35397	0.717	0.473338	
Tone3	1.61597	0.4525	3.571	0.000355	***
Tone4	0.18845	0.35186	0.536	0.592255	
Trainingvm	0.66776	0.50961	1.31	0.190081	
Trainingminimal	2.7766	0.73875	3.758	0.000171	***
Talkernew:Tone2	-0.40183	0.52142	-0.771	0.440912	
Talkernew:Tone3	-1.28277	0.60929	-2.105	0.035261	*
Talkernew:Tone4	-0.02918	0.53093	-0.055	0.956172	
Talkernew:Trainingvm	-0.44	0.55097	-0.799	0.424524	
Talkernew:Trainingminimal	-3.79612	0.7485	-5.072	3.94E-07	***
Tone2:Trainingvm	-0.08581	0.54113	-0.159	0.87401	
Tone3:Trainingvm	-2.18318	0.58986	-3.701	0.000215	***
Tone4:Trainingvm	-0.88103	0.51451	-1.712	0.086833	.
Tone2:Trainingminimal	-1.18514	0.80904	-1.465	0.142953	
Tone3:Trainingminimal	-1.61595	0.9646	-1.675	0.093885	.
Tone4:Trainingminimal	-2.81706	0.74397	-3.787	0.000153	***

Table D Logistic Regression Analysis of three training groups' word identification accuracy rate, using new talker stimuli, T4 and multi-talker training condition as the baselines.

	β	Std. Error	z value	P (2-tailed)	
(Intercept)	1.54005	0.377861	4.08E+00	4.59E-05	***
Talkerold	-0.57883	0.382153	-1.515	0.12986	
Tone1	-0.15928	0.397611	-0.401	0.68873	
Tone2	-0.3073	0.391246	-0.785	0.4322	
Tone3	0.173907	0.416118	0.418	0.676	
Trainingvm	-1.4286	0.506513	-2.82	0.0048	**
Trainingminimal	-1.41432	0.509223	-2.777	0.00548	**
Talkerold:Tone1	-0.02917	0.53093	-0.055	0.95618	
Talkerold:Tone2	0.372671	0.530931	0.702	0.48273	
Talkerold:Tone3	1.253597	0.617424	2.03	0.04232	*
Talkerold:Trainingvm	1.215319	0.511702	2.375	0.01755	*
Talkerold:Trainingminimal	1.373899	0.518994	2.647	0.00812	**
Tone1:Trainingvm	1.656345	0.549044	1.917	0.06255	
Tone2:Trainingvm	0.943807	0.518528	1.82	0.06873	
Tone3:Trainingvm	-2.28527	0.531728	-2.536	0.00161	**
Tone1:Trainingminimal	-3.34826	0.523451	-2.754	0.00068	**
Tone2:Trainingminimal	-0.74261	0.520809	1.506	0.13211	
Tone3:Trainingminimal	-2.17375	0.537313	-3.004	0.00647	**

Table E Logistic Regression Analysis of three training groups' word identification accuracy rate, using new talker stimuli, T4 and variance manipulated training condition as the baselines.

	β	Std. Error	z value	P (2-tailed)	
Talkerold	0.6365	0.34029	1.87	0.041418	*
Tone1	1.49707	0.37862	3.954	7.69E-05	***
Tone2	0.6365	0.34029	1.87	0.061419	.
Tone3	-0.11137	0.33103	-0.336	0.736551	
Trainingmultitalker	1.4286	0.50651	2.82	0.004795	**
Trainingminimal	0.01428	0.4799	0.03	0.976265	
Talkerold:Tone1	-0.80447	0.53236	-1.511	0.130749	
Talkerold:Tone2	1.22406	0.51368	2.301	0.021702	*
Talkerold:Tone3	0.23672	0.48312	0.49	0.624149	
Talkerold:Trainingmultitalker	-1.21533	0.5117	-2.375	0.017545	*
Talkerold:Trainingminimal	0.15857	0.48899	0.324	0.745726	
Tone1:Trainingmultitalker	-0.65635	0.54904	-1.195	0.255533	
Tone2:Trainingmultitalker	-0.9438	0.51853	-1.82	0.068735	.
Tone3:Trainingmultitalker	2.28528	0.53173	4.297	0.001607	**
Tone1:Trainingminimal	-1.26152	0.50918	-2.478	0.013228	*
Tone2:Trainingminimal	-0.15954	0.4837	-0.33	0.741535	
Tone3:Trainingminimal	0.28766	0.47448	0.606	0.544337	