NORTHWESTERN UNIVERSITY

Perception and Production of Non-Native Prosodic Categories

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Linguistics

By

Tessa Bent

EVANSTON, ILLINOIS

December 2005

ABSTRACT

Perception and Production of Non-Native Prosodic Categories

Tessa Bent

Cross-language phonological variation is manifested at the levels of segmental and prosodic structure. Modeling the interactions of two segmental systems during non-native speech perception and production has been highly successful (e.g. Best et al., 2001; Flege 1995; Kuhl and Iverson, 1995). However, these models lack an account for cross-language perception and production of prosodic structure. The perception and production of non-native prosodic contrasts was investigated by testing native English speakers' perception and production of Mandarin. Mandarin is a lexical tone language in which pitch changes over syllables signal differences in lexical meaning whereas in English pitch patterns over whole phrases express pragmatic meanings.

Native English and Mandarin listeners' perceptual sensitivity to Mandarin lexical tones was tested in monosyllabic and tri-syllabic utterances. The English listeners exhibited overall lower and more variable sensitivity compared to the Mandarin listeners. In particular, the English listeners' sensitivity to the Mandarin tone contrasts varied depending on the acoustic similarity of the pitch contours for contrasting tones and on whether the tones were presented in monosyllabic or tri-syllabic utterances. In contrast, the Mandarin listeners were highly sensitive to all contrasts. Furthermore, the English listeners appeared to have been influenced by assimilation to English prosodic categories. Finally, while Mandarin listeners mostly attended to lexical tone targets, English listeners attended mostly to global aspects of the stimuli.

In a tone production task, the English participants imitated native Mandarin tone productions. These imitations were then evaluated based on native Mandarin listener judgments and acoustic analysis. The English participants accurately imitated the contour shape but deviated from the Mandarin model in ways that were shifted towards English production norms. Listeners' perception and production abilities were not significantly correlated.

In sum, listeners may perceive non-native prosodic contrasts on the basis of a combination of language-independent auditory processing strategies and assimilation to native prosodic categories. Models of cross-language perception of prosody must incorporate contextual variation into their predictions and account for listeners' use of both auditory and linguistic processing. These studies serve as the basis for a long-term research program aimed at extending current models of cross-language speech perception to the prosodic level.

ACKNOWLEDGMENTS

I would first off like to thank my mom without whom none of this would have been possible. She has provided me with support, encouragement, love, and friendship. Throughout the most exciting achievements and the occasional disappointments, she has been the one person I knew I could always rely on to share in my joy or to offer words of encouragement in times of need.

My advisor, Ann Bradlow, introduced me to the field of phonetics and has showed me what an interesting and exciting field it is. I have appreciated Ann's clear teaching, committed mentoring, and willingness to give feedback. Her confidence in me has fueled my confidence in myself. Her tireless work to help me to improve each paper, presentation, grant proposal, and job application has been absolutely vital to my success.

My other committee members, Janet Pierrehumbert, Beverly Wright, and Bruce Smith have provided me with extremely helpful comments throughout the dissertation process. Their suggestions have helped to make my writing and presentations more clear and pushed my theoretical thinking further.

I would also like to acknowledge the contribution of two of my earlier (one much earlier) teachers: Linda Emerick and Kenneth Sheilds. Linda fostered a life long love of learning in me and Ken first introduced me to Linguistics.

Last but not least I would like to thank all my friends and family especially my grandparents, Geoff Edelmann, Jay Turner, Holly Boyle, Laura Brubaker, Caroline Campbell, Amy (Conn) Nohe, Terri Carr, Patty (O'Malley) Cardina, Diane Shearon,

Alise McClusky, Lyla Miller, Cara Coburn, Saundra Wright, Russ Burnett, Matt Fitzgerald, Elisa Sneed, James German, Ann Longawa, Rajka Smiljanic, Matt Goldrick, the Bunny Sharks, and Mingus.

This dissertation is dedicated to the memory of two people whom I truly wish were with me today to celebrate in my accomplishment: Robert Bigelow and Helen Gadsby.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

1. 1 Introduction

Cross-language phonological variation is manifested at the levels of segmental structure and suprasegmental structure. Segmental structure includes the consonant and vowel inventories of the language. Suprasegmental structure includes aspects of phonological structure larger than a segment including prosody, tone, and intonation. Prosodic categories convey both linguistic and paralinguistic information including lexical meaning, pragmatic meaning, emphasis, emotions, attitudes, and phrasing.

While much is known about how native segmental categories influence the perception and production of non-native segmental categories, very little is understood about the cross-language interaction of suprasegmental categories due to the fact that current models of cross-language speech perception focus exclusively on segmentals (Best, 1995; Best, McRoberts, and Goodell, 2001; Flege, 1986, 1995, 1999; Kuhl and Iverson, 1995; Kuhl, 2000). Prior research has established that non-native talker-listeners do not perceive or produce prosodic categories in the same way as native talker-listeners (Anderson-Hsieh et al., 1992; Chun, 1982; Derwing and Munro, 1997; Flege, 1995; Flege, Munro, and MacKay, 1995b; Kiriloff, 1969; Mennen, 2004; Scuffil, 1982). However, there is not a comprehensive explanation of how and why these differences are present. Whether the predictions of models of cross-language perception and production at the segmental level can be applied to suprasegmental categories remains an open question. The main question this dissertation addresses is, how do native language

prosodic categories influence the perception and production of a novel set of prosodic categories. That is, how does one deeply entrenched prosodic system influence a newly encountered system?

Languages vary widely in their realizations of the multiple dimensions included in the prosodic system. These dimensions include lexical prosody, intonation or post-lexical prosody, rhythmic structure, and the unit(s) which receive prosodic markers (Jun, 2005). This dissertation focuses on the dimensions of lexical prosody and intonation. The dimension of lexical prosody includes languages with either lexical tones, lexical stress, or pitch accent. The two languages under investigation, Mandarin and English, vary in their lexical prosody. Lexical tone languages, including Mandarin, have specified pitch patterns for each word. In lexical stress language like English, one syllable within each word will receive primary stress as marked by increased duration and amplitude. Furthermore, in English pitch accents can be associated with prominent words as determined by intonation. Intonation utilizes the dimensions of pitch, amplitude and duration to convey information about which words are stressed in the sentence and about the edges of prosodic units through the use of phrasal tones and/or boundary tones. All languages have intonation which interacts with the lexical prosody categories in the language. While some researchers only use the term prosody to refer to syllable structure, rhythm and phrasing and not intonation (Pierrehumbert, 1999), others include intonation (also called post-lexical prosody) within prosodic structure (Jun, 2005). Within this dissertation, prosody is used to refer to suprasegmental categories generally and will focus specifically on categories that use pitch as their primary acoustic correlate.

Whether the predictions of models of cross-language segmental perception can be brought to bear on cross-language perception of prosody first crucially depends on whether the system of native categories is used when interpreting non-native categories. Due to the wide typological variability in prosodic categories, it is not clear whether listeners will use their native prosodic categories to interpret non-native categories in the same way that non-native listeners have been found to use their native system of segmental categories to interpret non-native segmental systems. Whether non-native listeners use their native system of prosodic categories may depend on the degree of similarity between the native and non-native systems. One possibility is that listeners will not interpret a novel set of prosodic categories with reference to their native language prosodic categories. This possibility suggests that listeners will only use their psychoacoustic (i.e. language-independent) abilities when processing the non-native prosodic contrasts. In this case, all talker-listeners, regardless of the structure of their native language, should perform the same when attempting to perceive or produce a system of non-native prosodic categories. Another possibility is that non-native prosodic categories will be interpreted with reference to native prosodic categories and therefore, both language-specific and language-independent processing will be involved. For example, listeners from a lexical-stress language may use their pitch accent categories to interpret a lexical tone system. An intermediate possibility is that only listeners from similar types of prosodic languages will use their native categories to interpret non-native categories. Therefore, listeners from one tone language would interpret non-native tone categories with reference to their native set but listeners from lexical stress languages

would not perceive lexical tone categories in reference to their native language categories.

Recent accounts of the prosodic structure of many typologically and geographically diverse languages within the same theoretical framework (See all chapters in Jun, 2005) will assist with the comparisons of prosodic categories across languages and in the development of predictions about how any native prosodic system will influence the perception and production of any system of non-native prosodic categories. While a primary goal of cross-language research is to be able to account for the interaction of any two languages, most studies that have made concrete observations about transfer between prosodic structures in the native and non-native language have investigated languages that have typologically similar prosodic systems (e.g. Dutch and Greek (Mennen, 2004); English and German (Grover, Jamieson, and Dobrovolsky, 1987)). The goal of understanding when and how listeners use their native language to interpret the structures of a non-native language and to be able to predict how listeners from any one language will interpret any other is shared by the current investigation.

The perception and production of non-native prosodic categories were investigated through testing native Mandarin and English participants' discrimination and production of Mandarin tones in monosyllables and tri-syllabic sequences. The motivation for testing these two languages was to investigate the interaction of two very different prosodic systems in terms of the type of lexical prosody (tone versus stress) and the meanings conveyed by the prosodic categories (lexical and pragmatic versus just pragmatic). The long-term goal of this research is to extend current models of cross-

language speech perception to include the perception and production of non-native suprasegmental categories.

In the sections that follow, three current models of cross-language speech perception are described followed by discussion of the ways in which similarity of speech categories across languages are assessed. The next section explains the use of pitch in Mandarin and English and then a pilot experiment comparing the pitch contours in the two languages is presented. Lastly, previous research findings regarding the perception and production of Mandarin tones by native English speakers are described.

1.2 Models of Cross-Language Speech Perception

A comprehensive model of non-native sound structure acquisition must account for two key observations. First, a model must account for the decline over the life span in humans' ability to perceive and produce non-native speech sounds in a native way (Long, 1990; Flege, 1995; Flege, Munro, and MacKay, 1995a). Second, a model must account for the variable ease of acquisition of non-native sounds. Several models of non-native phoneme perception have been proposed to explain why the perception and production of many non-native phoneme contrasts are difficult for adults: the Perceptual Assimilation Model (PAM) (Best, McRoberts and Sithole, 1988; Best, 1995; Best, McRoberts, and Goodell, 2001), the Speech Learning Model (SLM) (Flege, 1986, 1995, 1999), and the Native Language Magnet Model (NLM) (Grieser and Kuhl, 1989; Iverson and Kuhl 1996; Kuhl, 2000; Kuhl and Iverson, 1995).

All the models agree that listeners do not perceive non-native sounds in a native-like way because they already have a native language system in place. Adult learners tend to interpret novel sounds with reference to the categories of their native language system. Furthermore, they have learned to attend to dimensions that are important for distinguishing categories in their native language and to "ignore" within-category variation under typical listening conditions. Therefore, the differences between the categories in the native language and the non-native language can cause difficulties for non-native speech perception and production. Although there is general agreement among the three models on these points, the focus and theoretical underpinnings of each model are different.

1.2.1 The Perceptual Assimilation Model

The Perceptual Assimilation Model (PAM) focuses on the abilities of naïve non-native listeners to perceptually discriminate pairs of non-native phonemes. The PAM hypothesizes that listeners assimilate non-native sounds to the closest sound in their native system in terms of articulatory similarity. Predictions of discrimination patterns are based on the relationship between the two sounds to be discriminated and their assimilation to native categories. In this model, non-native phonemes can be assimilated to native categories in one of the following three ways: (1) assimilated to a native category – the fit can vary from good to notably deviant, (2) assimilated as an uncategorizable speech sound – that is, the phoneme is perceived as speech but cannot be assigned to any native category, or (3) nonassimilable non-speech sound – that is, sounds

that are not assimilated into the native phonological space. Within these three assimilation categories, the types of knowledge that listeners will bring to bear on the task will differ. That is, if a sound is assimilated to a native category, then primarily linguistic-phonetic knowledge will be recruited. For sounds that are assimilated as uncategoriezable speech sounds, listeners will process the sounds using a combination of linguistic and auditory processing. For nonassimilable non-speech sounds, listeners will process the sounds in an auditory only mode.

PAM predicts patterns of discrimination depending on the fit of the members of the pair into the three assimilation categories. The following six categories describe patterns of pair-wise assimilation and the associated predicted patterns of discriminability: (1) A Two-Category Assimilation is an instance where each member of the pair is assimilated to a different native category. Discrimination is expected to be excellent; (2) A Single-Category Assimilation will result when both non-native phonemes assimilate to the same native phoneme category. Discrimination is predicted to be poor; (3) A Category-Goodness Difference will result when both non-native phonemes assimilate to the same native category but one fits better than the other. Discrimination is expected to be moderate to very good; (4) A Categorized-Uncategorized Pair will result when one non-native phoneme assimilates to a native category and one cannot be assimilated to any native category. Discrimination is expected to be very good; (5) A Both Uncategorizable pair will result when both non-native speech sounds are perceived as speech but cannot be assimilated to any native speech categories. Discrimination is expected to be poor to very good depending on the similarity of the two sounds to one

another and to other native phonemes; (6) A Nonassimilable pair results when both of the speech sounds are perceive as non-speech sounds. In this case, discrimination should be good to very good. PAM's predictions have been tested using a wide variety of languages and phoneme contrast types including English listeners perception of Zulu clicks (Best et al., 1988; Best et al., 2001), Farsi consonants (Polka, 1992), Salish consonants (Polka, 1992), Hindi stop consonants (Polka, 1991) and Norwegian, French, and Thai vowels (Best, Faber, and Levitt, 1996); Japanese and French listeners perception of English approximants (Best and Strange, 1992; Halle, Chang, and Best, 1999); and Japanese listeners perception of a variety of English consonants (Guion et al., 2000). Generally, the predictions of the model have been borne out. However, the vast majority of the work testing the model has used consonants and isolated monosyllables and the predictions are based on context-independent phonemes without an account of allophonic variation across languages. Therefore, the generalizability of these predictions to sounds in running speech is still undetermined and vowel categories remain understudied. Furthermore, while the model assumes that the representations for perception and production are shared, predictions are not made about the production of non-native segmental categories.

Halle et al. (2004) represents the only attempt, thus far, to apply the predictions of PAM to prosodic contrasts. In this study, French and Taiwanese listeners identified and discriminated three tone continua that were derived from naturally produced Mandarin Chinese tone contrasts. Taiwanese listeners were found to exhibit quasi-categorical perception of the tones whereas French listeners seemed to be relying more on

psychophysical factors and perceived the tones as nonlinguistic melodic variations. French listeners' performance, while worse than the Taiwanese listeners, was not poor. Halle et al. speculate that the French listeners may be sensitive to these tones due to their experience with intonation contours meant to convey emotion and may map certain tones onto these paralinguistic uses of intonation (e.g. forbidding or stunned) although they state that these are "loosely defined" and are not a set of finite contrastive categories. They speculate that the mapping from French to Mandarin may be one of a Both Uncategorizable pair where discrimination would depend on perceived similarity between the contours and their relationship to native language contours.

1.2.2 The Speech Learning Model

The Speech Learning Model (SLM) focuses on second language learning and the connection between perception and production. According to the SLM, the continued use of the first language and the different contexts and nature of input during second language acquisition are the factors that constrain an adult's ability to fully acquire a non-native language. However, the ability to acquire new representations for speech sounds remain intact across the life span, but it becomes more difficult as the age of first exposure increases. The SLM proposes that when the native language is firmly established, as is the case for adult learners of a second language, non-natives are more likely to interpret novel sounds with respect to the native system and therefore will have difficulty establishing the necessary new categories in the second language.

The SLM hypothesizes a process of assimilation where non-natives assimilate a novel segment into a native category in order to form a composite category. This process of assimilation is similar to the assimilation to a native category process in the PAM model. However, in the PAM, predictions are only made about naïve perception whereas the SLM focuses on long term representation of native and non-native phoneme categories. In the SLM, a process of dissimilation between native and non-native segmental categories can also occur resulting in two categories, which are adjacent to one another in the shared phonetic space, moving further from the native language norms in each language. Both of these processes can result in non-native categories that range from good to deviant exemplars of the native category. An important feature of this model is that it is explicit about the bi-directional nature of the interaction between the native and non-native languages in an individual. The SLM is also explicit about the perception-production relationship. In the SLM, perception at first leads production and then as learners gain proficiency their perception and production of non-native sounds are brought into closer correspondence with each other. The predictions of the SLM have been tested using both consonants (Flege, 1987; Flege and Eefting, 1987; MacKay et al., 2001) and vowels (Flege, MacKay and Meador, 1999; Flege, Schirru, and MacKay, 2003; Flege and MacKay, 2004). These experiments have generally upheld the predictions in the model (however, see Guion et al., 2000 for some evidence against the predictions with relatively inexperienced second language learners).

1.2.3 The Native Language Magnet Model

The Native Language Magnet Model (NLM) focuses on within-category structure for phoneme perception and assumes that speech perception is processed using a general auditory-acoustic mechanism. In the NLM, exposure to the native language alters the perceived distances between category exemplars in the acoustic space and these alterations lead to long-term changes in speech perception patterns. Phonetic prototypes are formed through exposure to the distribution of sounds encountered in the native language. The NLM is not explicit as to whether these phoneme prototypes are stored as abstract summary representations or as the most representative stimulus within a set of exemplars. Sounds close to the prototype are perceptually drawn to it, and therefore, the perceived distance between the prototype and other members of the category shrinks. This alteration in perceptual space can enhance discrimination performance for non-native sounds by heightening the relative salience of essential acoustic cues if these cues are also important for making between-category distinctions in the native language. However, this alteration can hinder discrimination by attenuating cues that signal within-category variation in the native language, but are important for making between-category distinctions in the non-native language. Evidence in support of the NLM model has come from studies of vowels (Kuhl, 1991; Iverson and Kuhl, 1995) and sonorants (Iverson et al., 2003).

1.3 Assessing Similarity

While current models of cross-language speech perception depend on the ability to assess similarity of phoneme categories across languages, the appropriate way to assess similarity has been a subject of recent research (e.g. Strange et al., 2004). Strange and her colleagues stress the importance of establishing ways of determining cross-language perceptual similarity independent of identification or discrimination tests in order to predict initial perception performance of non-native contrasts and patterns of difficulty in L2 learning. Previous methods of assessing similarity have included acoustic comparisons across languages (Strange et al., 2004), comparison of articulatory-phonetic characteristics across categories (Best et al. 2001), direct assessments of similarity by explicitly asking listeners to place non-native segments into native categories (Guion, et al., 2000; Iverson et al., 2003; Strange et al., 2004), listener transcriptions of non-native sounds using native spelling and other descriptions (Best et al., 2001), and judgments of non-native phonemes' fit into native categories (Guion et al., 2000; Iverson et al., 2003). Comparisons of these methods and their predictive value are important in advancing the models of cross-language speech perception. The most appropriate method may differ depending on the type of category (i.e. vowels or consonants) and needs to take into account both correspondences across abstract language categories and talker and listener variability.

The picture of assessing similarity for segmental categories across languages has been further complicated by observations that speech categories vary acoustically depending on the segmental and prosodic context. This variation affects perceptual

assimilation judgments, which may or may not be predictable from the acoustic variation caused by the contextual variation. Both Best et al. (2001) and Strange et al. (2004) note the need for assessing similarity in longer contexts than just a syllable to better approximate natural running speech and to take into account such factors as cross-language differences in phonotactic constraints and coarticulatory patterns. Identification and discrimination accuracy also depends on the context in which the segments are placed. For example, the widely cited case of Japanese listeners' difficulty with English /r/ and /l/ actually varies significantly depending on the placement of these phonemes in a word, with word-initial position being difficult, especially within consonant clusters, and word-final position being relatively easy (Logan, Lively, and Pisoni, 1991). Strange et al. (2001) found that the perceived similarity between English and Japanese vowel categories varied depending on the speaker and the consonantal context in which they were presented. From these observations, it is clear that comparing abstract phoneme categories will not be sufficient to fully describe how non-native phonemic categories are assimilated. Taking into account the fine phonetic details influenced by the surrounding segmental and prosodic context will be essential in increasing the predictive value of current models of cross-language speech perception.

Many of the same unresolved issues of comparing segmental categories across languages are also at issue for suprasegmental categories. Some additional difficulties also exist for suprasegmental categories. Both of these types of issues are discussed below. First, the contextual variation just discussed for segmental categories also influences the realization of suprasegmental categories (e.g. Ladd and Schepman, 2003;

Pierrehumbert and Hirschberg, 1990; Xu, 1997). Suprasegmental categories vary widely depending on the segmental material over which they are placed, the interaction of different levels of suprasegmental structure, and the surrounding suprasegmental material. This variation needs to be taken into account when comparing suprasegmental structures across languages. Second, the identification of the relevant segmental comparisons across the languages is typically straightforward but may be more complicated for suprasegmentals. When investigating phoneme perception in a non-native language only the phonemes or a subset of the phonemes in the language need to be contrasted. For example, when studying Japanese listeners perception of English, a relevant set of phonemes may be identified such as English /r/ and /l/ and Japanese /r/ and /w/. However, for prosodic categories more than one level of the sound system may need to be investigated as a source of possible transfer and interference including lexical tones, word stress, pitch accent, and sentence or phrasal intonation patterns. Third, in investigations of phoneme perception and production, the perceptually relevant acoustic dimensions can usually be determined without explicit investigation. For the previous example, before the experiment began it could easily be determined that the most relevant acoustic dimension is the third formant, although the other formants and duration may also play a role. However, with prosodic categories, it may be more difficult to determine the relevant acoustic dimensions. Pitch height and the shape of the pitch contour will certainly be important but duration, amplitude contour, and voice quality can also contribute to category assignment. Lastly, the labels for the phonemic categories are typically familiar or can be easily taught to participants. With the identification of /r/ and

/l/, although the Japanese participants may not be able to identify these two consonants, most participants would be familiar with the labels. However, with suprasegmentals even native speakers may not have overt labels for the categories or the system of labels may not be rich enough to account for the observed number of patterns. For example, native English speakers may have labels for several different intonation patterns such as statement, command, or question, but naïve listeners do not have as rich a labeling system as exists in English and has been described by researchers (e.g., in the ToBi system) (de Bot, 1980).

The methods used for assessing similarity for vowels and consonants have generally not been tested for suprasegmental categories except for a few studies which have used impressionistic judgments (e.g. Chen, 1997; Halle et al., 2003; White, 1981) and acoustic comparisons (Grover, Jamieson, and Dobrovolsky, 1987; Mennen, 2004). Furthermore, some of the methods cannot be used with suprasegmental features such as having listeners describe which native category a non-native feature best fits into or transcribing non-native sounds using native orthography. Listeners limited metalinguistic awareness of suprasegmental features and the lack of encoding of many suprasegmental features in the native language orthography severely limits the wholesale adoption of the segmental similarity assessment techniques for suprasegmentals. Therefore, research must be conducted to determine which categories in the native language affect perception and production of non-native suprasegmental categories and how these categories influence one another for non-native listeners. This research must use a protocol that allows for these types of investigations without requiring participants to have overt labels

for the native and non-native suprasegmental categories. Further, these cross-language comparisons must take into account contextual influences.

One technique for assessing the perceptual similarity of categories, including speech categories, is multidimensional scaling (MDS). Because this technique requires a procedure in which listeners make pair-wise comparisons (e.g. same-different judgements or magnitude estimation of similarity), this technique is particularly useful in cases where there may not be well-known, conventionalized labels for the categories under study as is the case with suprasegmental categories. Furthermore, based on the derived similarity spaces for the stimulus items, the features that participants are using to base their judgments of similarity can be inferred. Within speech research, MDS has been used to determine similarity relationships and feature perception for a variety of stimuli including consonants (Iverson and Kuhl, 1996; Iverson et al., 2003), lexical tones (Gandour and Harshman, 1978; Huang, 2001; 2004), intonation contours (Grabe et al., 2003), and vowels (Fox, Flege, and Munro, 1995; Iverson and Kuhl, 1995; Kewley-Port and Atal, 1989).

In a seminal study of lexical tone perception, Gandour and Harshman (1978) attempted to determine the number and nature of features used during tone perception and to determine how linguistic experience shapes the way that these features are exploited. In their study, Thai, Yoruba, and American English listeners rated the similarity of a set of thirteen synthesized tones which differed in terms of beginning and ending pitch height, contour versus level, slope magnitude, pitch range, and duration. The results of their MDS analysis indicated that the listeners used the following

dimensions to rate the stimuli: (1) average pitch, (2) direction, (3) length, (4) extreme end point, and (5) slope. Gandour and Harshman suggest that some of these dimensions are clearly linguistic-phonetic (direction and slope) while others are either potentially non-linguistic/auditory (average pitch and length) or definitely non-linguistic/auditory (extreme end point). Although all three listener groups used all five dimensions, the weight that they attached to these dimensions varied. Hombert (1976) also used MDS to determine the features that Yoruba listeners were using in their similarity judgments of disyllabic noun pairs. These listeners attended to two of the same dimension as in Gandour and Harshman's study: direction of pitch movement and slope. More recently Huang (2004) has investigated tone perception by native Chinese and English listeners. In her experiments, English listeners attended primarily to pitch onset and offset (which she claims are psychoacoustically based) whereas Chinese listeners paid attention to contours as a whole and were influenced by the tone sandhi rules in their particular dialect. A similar technique has been used to test English, Spanish and Chinese listeners' perception of English intonation contours (Grabe et al., 2003). The direction of change was the most important feature for all the listeners suggesting a universal auditory mechanism, but the listeners' arrangement of contours in perceptual space was warped by linguistic experience. Together, these MDS studies suggest that listeners process non-native prosodic categories in a primarily auditory mode although linguistic-phonetic processing may be secondarily activated.

1.4 Pitch in Mandarin and English

Both Mandarin and English utilize pitch variation to convey prominence and pragmatic meaning while only Mandarin encodes lexical meaning through pitch cues. The broad patterns of pitch variation are similar in the two languages but details of timing and alignment as well as segmental material over which the contours are placed differ substantially. Below is a description of how pitch is used in these two language.

1.4.1 Pitch in Mandarin: Lexical Tones and Intonation

Mandarin, a tone language, has four different pitch patterns (tones) that can lexically distinguish words with the same phonemes: Tone 1 has a high-steady pitch, Tone 2 a high-rising pitch, Tone 3 a low-dipping pitch, and Tone 4 a high-falling pitch (Chao, 1968). The tones can also be described with reference to a five-point pitch height scale (5 = high, 3 = mid, and 1 = low) in which the two pitch height numbers for each tone indicate the values at the beginning and end of the syllable. In this system, Tone 1 is 55, Tone 2 is 35, Tone 3 is 21(3) and Tone 4 is 51. There is also a neutral tone that has a very small pitch range and is short in duration which is used mostly for particles and unstressed syllables in words. The pitch of the preceding syllable determines the pitch height of the neutral tone. Mandarin listeners are easily able to identify the tones of both real-speech isolated monosyllabic words (Chuang et al., 1972) and non-words (Bent, Bradlow, and Wright, in press). When Mandarin listeners make lexical tone identification errors, they tend to confuse Tone 2 and Tone 3.

Many contextual effects influence the realization of tonal pitch patterns. These effects stem from either voluntary (i.e. linguistic/paralinguistic factors) or involuntary (i.e. articulatory constraints) factors (Xu, 2001). Additionally, tones produced in sequence are influenced by both the tones of the preceding and following syllables and by tone sandhi rules. While both anticipatory and carry-over tonal coarticulation influence tone realization, the effect of carry-over articulation tends to be much larger than that of anticipatory coarticulation (Xu, 1997). The low tone (Tone 3) is changed most from its citation form to its form in context by two additional factors. First, the low tone loses its rise at the end so that in isolation or at the end of an utterance the low tone can be described as a 213 contour but in context it is typically 21. Also, due to tone sandhi, the low tone changes to the rising tone when followed by another low tone. While previous research has suggested that the difference between Tone 2 and Tone 3 is neutralized in the tone sandhi environment (Wang and Li, 1967), there is question as to whether this categorical tone sandhi effect takes place in normal running speech (Kratochvil, 1987). Kratochvil notes that the application of tone sandhi in running speech is rare and "appears to be limited to recurrent strings, the resulting shape is not that of Tone 2 but an idiosyncratic form distinct from both Tone 3 and Tone 2" (p. 425). However, in an identification task using disyllables with underlying forms of Tones 3-3 or Tones 2-3, Mandarin listeners were not able to correctly identify the first tone in the sequence (Wang and Li, 1967) and the acoustic differences between the underlying Tones 2 and 3 are extremely small (Peng, 1996). There are two other tone sandhi rules that may be considered phonetic rather than phonological. Chao (1965, 1968) describes these rules as

Tone 2 becoming Tone 4 when it is preceded by either Tones 2 or 4 and followed by a tone of any value. The other rule is Tone 1 changing to Tone 2 before Tone 4. This rule is not obligatory, however.

Lexical tone is not the only factor that influences the pitch of Mandarin utterances. Larger intonation units also influence the realization of the tones. For example, tone contours are higher in questions than in statements (Shen, 1990). Stress influences tone realization as well: a stressed syllable's pitch range is relatively widened and the slopes are more pronounced (Kratochvil, 1998), whereas a completely unstressed syllable is reduced to the neutral tone. Kratochvil (1998) also notes that in the Beijing dialect there is a tendency towards an alternating stress pattern of the iambic type.

In addition to the differences in fundamental frequency that distinguish the four tones, there are other differences among the tones which may further enhance the four way contrast including duration, amplitude envelope, and voice quality. Some of these acoustic differences between tones may be due to their auditory enhancement value and are not simply associations between two features that are learned by native language speakers (Blicher et al., 1990). For both native Mandarin and English listeners, a longer stimulus duration shifted identification responses for Tones 2 and 3 responses towards more Tone 3 responses. Blicher et al.'s interpretation of these results was that the lengthening of the stimuli made the initial falling portion of Tone 3 more detectable and therefore more distinct from Tone 2. Results from Moore and Jongman (1997) support this result as they show that stimuli with later turning points and larger falls in fundamental frequency will shift responses toward Tone 3. Fu et al. (1998) also note the

importance of temporal waveform envelope cues in the identification of tone, especially for Tones 3 and 4 since the amplitude envelope and F0 contour are highly correlated. Tone 3 is especially differentiated by the shape of its amplitude contour since the contour tends to have a falling-rising shape in comparison to the other three tones which have amplitude contours with rising-falling shapes.

1.4.2 Pitch in English Intonation

The way pitch is used in English differs from Mandarin in several ways. First, in English pitch is used systematically at the phrasal level rather than at the syllable level. Second, English pitch patterns do not express lexical meaning as in Mandarin but rather pragmatic meaning. For example, the sentence "Legumes are a good source of vitamins" can either be interpreted as a statement if there is a falling-rising pattern on *vitamins* or can be interpreted as a question if there is a rising pattern on *vitamins* (Pierrehumbert, 1980). Furthermore, there is some evidence that Mandarin is characterized by a greater fundamental frequency range than English (Chen, 1972; Chen, 2005).

Pierrehumbert and Hirschberg (1990) identify four important features in English intonation patterns: stress, tune, phrasing, and pitch range. Tune is perhaps the most relevant of these dimensions when comparing the English pitch patterns to Mandarin. The tune is "the abstract source of fundamental frequency patterns" (p. 272) and is made up of pitch accent(s), phrase accent(s) and the boundary tone. The different pitch accents mark the prominence of lexical items (Beckman and Pierrehumbert, 1986). There are five pitch accents, which can either include one or two tonal targets. These pitch accents

are described with an L for a low tonal target, an H for a high tonal target, and a star to denote alignment of the tone with the accented syllable. The exclamation point diacritic represents a "downstepped" tone or a compression of the pitch range. The pitch accents are H*, L*, L+H*, L*+H, H+!H*. The phrase accent (either high, H-, or low, L-) controls the pitch between the last pitch accent and the end of the phrase. The boundary tone (either high, H%, or low, L%) is the final tone, which marks the end of an intonation phrase. These pitch accent(s), phrase accent(s) and the boundary tone can be combined to produce at least 22 different contours. Within these contours there are pitch patterns that are broadly comparable to those seen on Mandarin syllables (i.e. level, rising, falling, dipping). These types of contours are contrastive and result in differences in pragmatic meaning.

1.4.3 A Comparison of English and Mandarin Pitch Contours

In order to test the hypotheses in current models of cross-language speech perception, it is essential to be able to assess the similarity of the speech categories in the two languages under study. Towards this goal, the similarity of pitch contours in English and Mandarin were compared. As an initial attempt at making comparisons across the two languages, all the possible Mandarin tonal combinations in tri-syllabic sequences were recorded. These contours were then mapped onto the prosodic contours in English created by the different combinations of pitch accent, phrase accent, and boundary tone as presented in Pierrehumbert and Hirschberg (1990). This impressionistic attempt was undertaken because other methods of assessing similarity such as listener transcriptions,

having listeners place non-native productions into native categories, and judgments of fit or goodness could not be applied to Mandarin tonal contours due to naïve listeners' lack of sufficient labels for English intonation contours.

For this informal experiment, a list of all the possible tonal combinations of two and three syllable utterances was constructed as well as all four tones in isolation and a few other select combinations of tones that might approximate English intonation contours (See Appendix A for the list of stimulus items). All of these stimuli were real words in Mandarin and used all voiced segments to get an uninterrupted pitch contour. One female native Mandarin speaker was recorded reading the stimuli. The recordings were made on an Ariel Proport A/D soundboard with a Shure SM81 microphone in a sound-attenuated booth. After the recording, sound files were converted to the WAV format with a 22.05 kHz sampling rate and transferred to a PC-based computer. The digital speech files were then segmented into individual stimulus files. These recordings were then played to an expert in English intonation (Janet Pierrehumbert) to determine whether these contours approximated their closest prosodic contour counterparts in English. Although Pierrehumbert noted some similarities in parts of these Mandarin tonal contours, she did not think that any of the contours closely approximated English intonation contours. Listening to these contours, one would certainly not mistake them for English. Furthermore, there were many contours that could not be matched with English categories as well as many English prosodic contours that did not seem to have any counterpart in Mandarin, in particular all the contours with a level tone in the middle of the pitch range.

Four examples of the "matches" between English intonation contours and my recorded Mandarin contours are shown below. The Mandarin examples shown below are taken from recordings of talkers used in the perception experiments reported later in this dissertation. The English contours are taken from a set of ToBi training materials (Beckman and Elam, 1997).

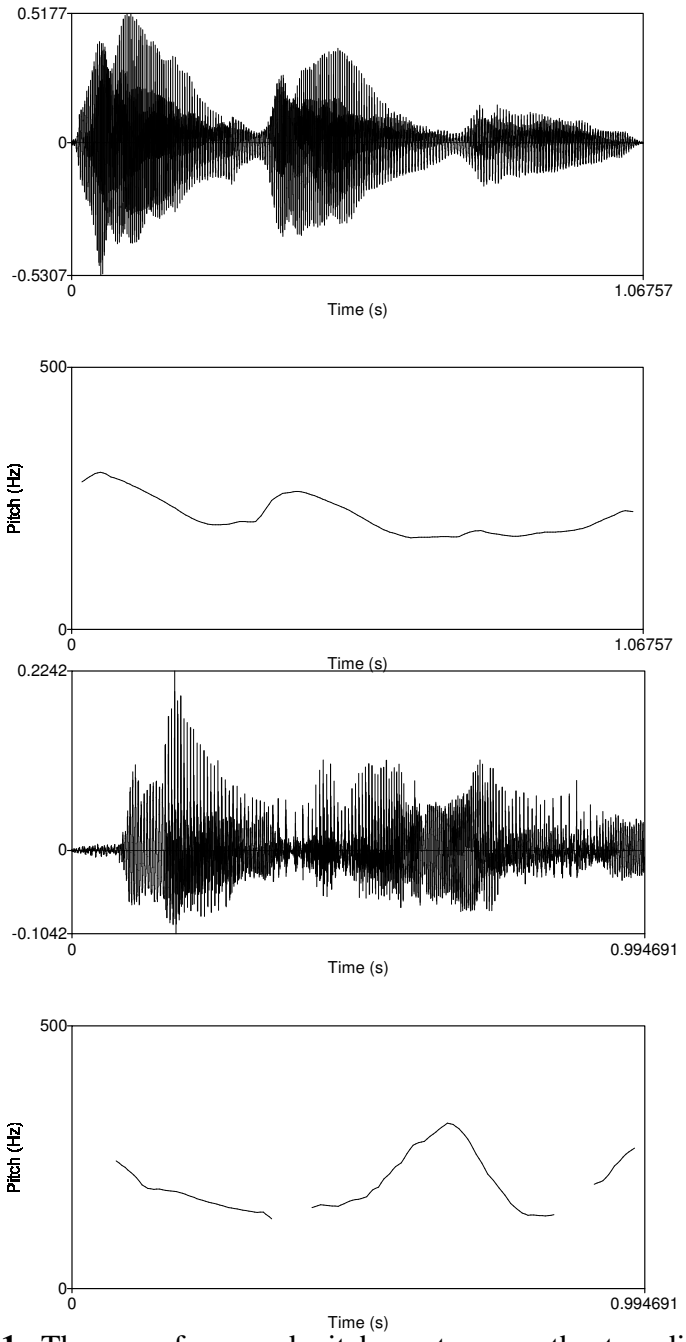


Figure 1.1: The waveform and pitch contour on the top display a female Mandarin speaker's utterance /ra ra ra/ with the falling tone on the first two syllables and the rising tone on the third syllable. The waveform and pitch contour on the bottom display a native English speaker producing "Marianna" excised from the sentence, "Marianna made the

marmalade." The intonation contour on Marianna is made up of the following pitch accent, phrase accent and boundary tone: L+H* L- H%.

Figure 1.1 demonstrates that although the general shapes of the contours in the two languages are the same (in this case a pattern of falling, rising, falling, rising), the placement of the pitch peaks and valleys can vary dramatically. These two contours are essentially the same duration (three syllables and 1067 ms for the Mandarin utterance and four syllables and 995 ms for the English utterance). For Mandarin, this general placement of peaks for a series of two falling tones followed by a rising tone will be basically the same for all utterance that have this series of tones. However, the placement of the peak for this English intonation contour will depend on the placement of stress within the word or phrase. Additionally, the timing between the pitch events can vary dramatically as demonstrated in the next example.

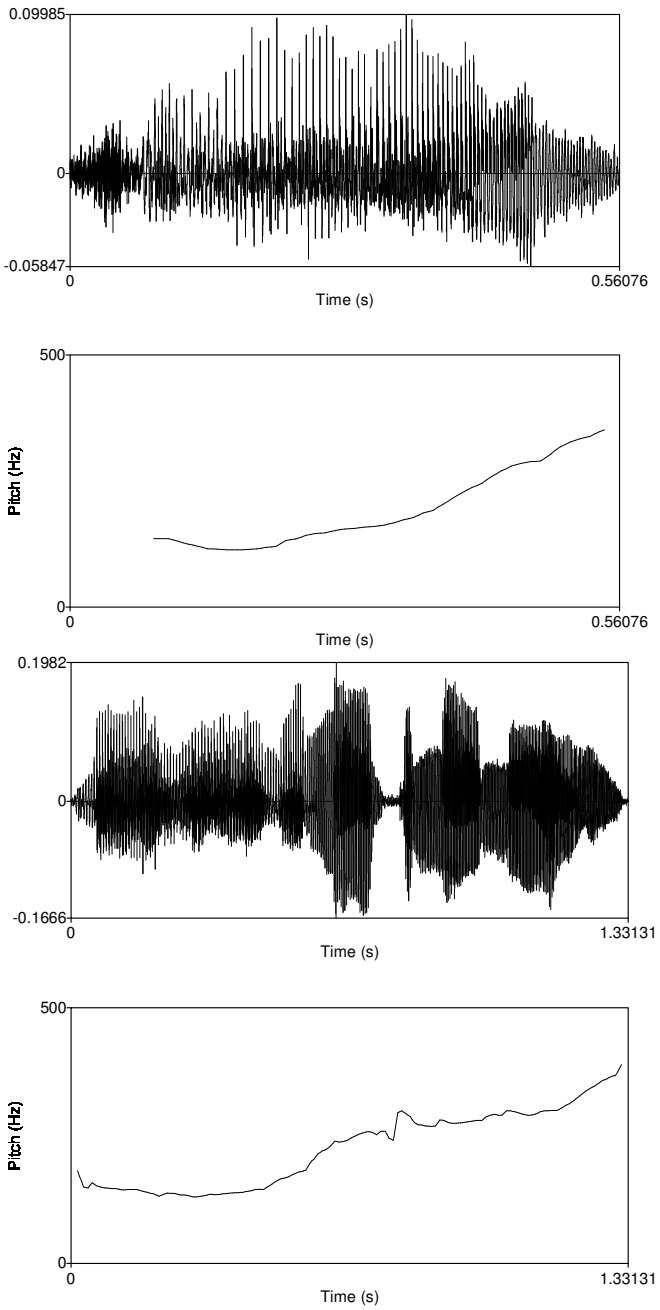


Figure 1.2: Both contours show a female English speaker producing an L* H-H% contour. The utterance on the top is "jam" excised from "Will you have marmalade, or jam?" and the utterance on the bottom is "Marianna made the marmalade."

Figure 1.2 demonstrates that in English the same intonation contour can be draped over vastly different amounts of segmental material. The example here is a one-syllable word versus a nine-syllable sentence. The contour shape is generally the same, however. The contour shape is also very similar to the rising tone in Mandarin as shown in Figure 1.3 below. In contrast to English, the Mandarin rising contour will always be applied to a single syllable. There are no patterns like this one that would arise on a multi-syllabic phrase in Mandarin.

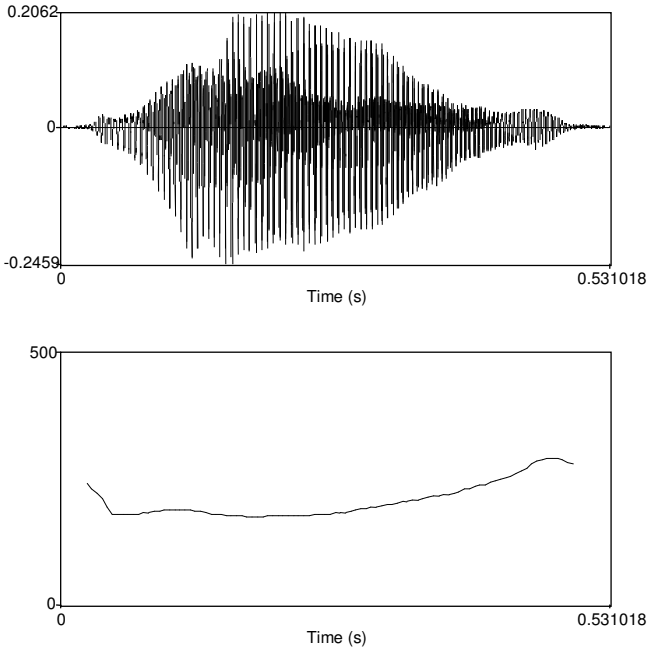


Figure 1.3: The production of "ra" with the rising tone by a female speaker of Mandarin.

Another difference between contours in Mandarin and English is pitch height. While the contour shape may be similar in both languages, their pitch heights may differ. One example is shown in Figure 1.4 in which level contours in English and Mandarin are

illustrated. In English, there is a mid-level contour (shown on the left), which does not exist in Mandarin. In Mandarin, the high level tone, while similar in contour shape to the one seen in English, is not common in English, as even when there are several high targets in a row, there tends to be more pitch movement. For example, an H* pitch accent followed by a H-H% will result in a high rising contour rather than a high level contour.

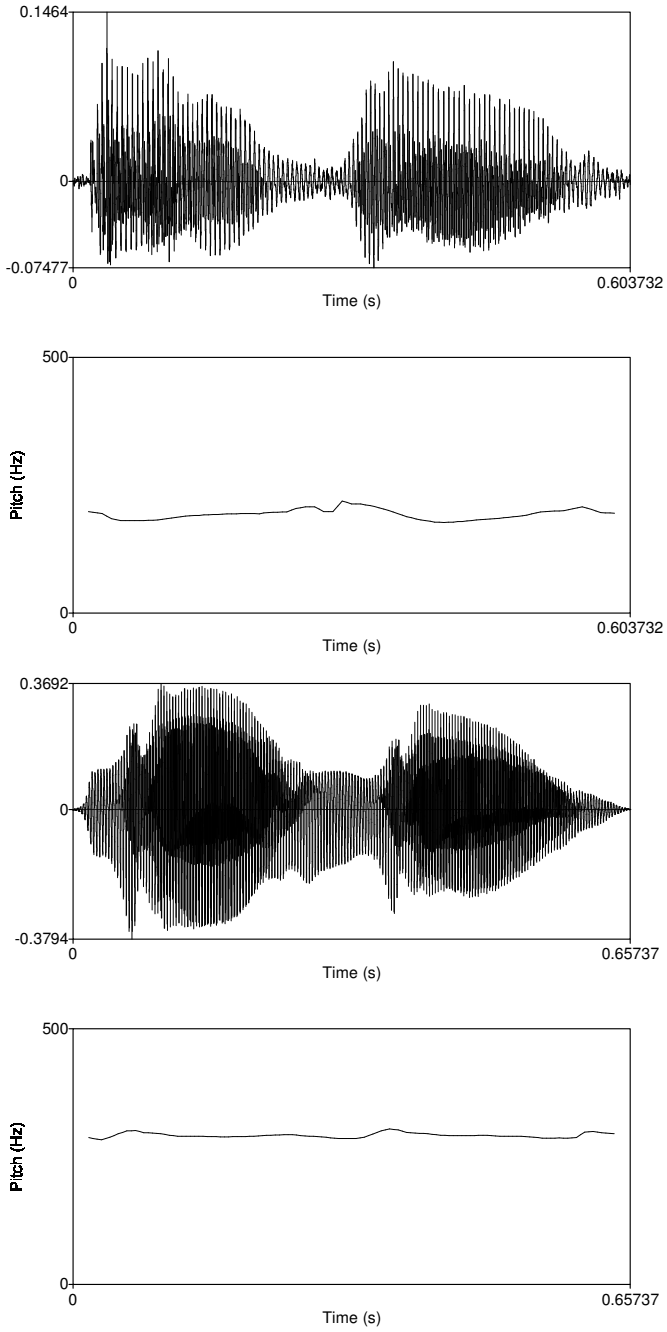


Figure 1.4: The waveform and pitch contour on the top show the production by a female English speaker of "basil" extracted from the sentence, "It's got some oregano 'n marjoram 'n some fresh basil." The contour on the bottom shows a female Mandarin speaker producing two high level tones on the syllable "ra". The English speaker's

production is approximately in the middle of her pitch range while the Mandarin speaker's is in her upper pitch range.

The above examples illustrate that while many of the types of pitch contours in English and Mandarin are broadly comparable, there are many differences in the details of their realization including factors of pitch range, alignment with stressed syllables, and amount of syllabic material over which the contour can be placed.

1.4.4 Perception and Production of Mandarin Tone Categories by Native English Listeners

The identification and discrimination of Mandarin tone categories in monosyllables is difficult for native English listeners (e.g., Kiriloff, 1969; Shen, 1989; Wang et al., 1999; Gottfried and Suiter, 1997; Lee et al. 1996). Furthermore, learning of tone for adults is more difficult than learning new segments while the reverse is true for children (Ioup and Tensomboon, 1987). This attenuation in the ability to discriminate tones that are not encoded in a listener's language is seen by 9 months of age (Mattock and Burnham, 2003). Various explanations have been put forth to explain this difficulty. An overly simplified explanation is that native English listeners are moving from a one-category system (no differences in lexical tone) to a four-category system. Therefore, syllables that only differ in lexical tone would fall into one category and discrimination will be difficult for all tones as long as they are perceived as speech (Burnham et al., 1996). However, this explanation does not account for differences in identification

accuracy, confusability, and rate of learning among the tones (Kiriloff, 1969; Wang et al., 1999). Another proposal, which can account for some of the perceptual differences among the tones, relates the Mandarin tone categories to the pitch patterns in English. The following factors have been proposed as contributing to the ease or difficulty in perceiving and producing certain tones: stress patterns (White, 1981; Shen, 1989), unnaturalness or novelty (Wang et al., 1999), and pragmatic meaning (Chen, 1997; White, 1981). However, a theoretically informed comparison of the patterns in the two languages is lacking. Generally, interference from English phonological structure is considered as a hindrance to accurate tonal perception and production. A non-language-specific factor that has been proposed to account for patterns of perceptual confusion is the psychoacoustic similarity of pairs of tones (Chen, 1997; Halle et al., 2004; Moore and Jongman, 1997). Generally the rising (35) and dipping (21(3)) tones have been found to be the most confusable for English listeners (Blicher, Diehl, and Cohen, 1990; Gottfried and Suiter, 1997; Leather, 1990; Shen and Lin, 1991; Whalen and Xu, 1992).

While there is no definitive explanation for the difficulty that English listeners have with Mandarin tones, there are many differences in the way that English listeners perceive Mandarin tones compared with native Mandarin listeners. First, the perceptual salience of acoustic cues seems to differ between the listeners. Previous studies have shown that both with vowels (Iverson et al., 2003) and tones (Gandour and Harshman, 1978), linguistic experience influences the weight that listeners place on various cues and can result in inaccurate identification of phonemic categories. Huang (2004) investigated the discrimination of Mandarin isolated monosyllabic tones by English and Mandarin

listeners and found that the listener groups were not attending to the same aspects of the stimuli. While both Mandarin and English listeners attended to starting pitch height, the Mandarin listeners placed importance on whether the tone was static or dynamic whereas the English listeners attended to ending pitch height. Second, discrimination experiments with Mandarin and English listeners have demonstrated that the discrimination functions are different for the groups of listeners both in terms of the shape of the function and the location of the boundary (Chan, Chuang, and Wang, 1975; Halle et al., 2004; Wang, 1976). Third, while surrounding context seems to help Mandarin listeners identify tones, native English listeners are hindered by the presence of a following syllable (Gottfried and Suiter, 1997). Fourth, native Mandarin listeners are better able to take advantage of information other than the pitch contour (like creaky voice) than English listeners when the information presented to them is reduced. Mandarin listeners are overall less dependent on redundant information presented throughout the syllable (Gottfried and Suiter, 1997).

In addition to the differences observed in behavioral tasks, neurophysiological methods have also shown differences between English and Mandarin listeners when they are presented with Mandarin tones. Specifically, native Mandarin and English listeners seem to employ different parts of their brains when listening to Mandarin tones. Research using both dichotic listening tasks (Wang, Jongman, and Sereno, 2001) and Positron Emission Tomography (Gandour, Wong, and Hutchins, 1998; Gandour et al., 2000; Wong et al., 2004) have found that tone language listeners tend to use the left side of their brain when processing tones from their own language while non-tone listeners (or

non-native speakers of that particular tone language) use either their right side or both hemispheres. Therefore, the lateralization of pitch processing depends both upon the linguistic use of the pitch contrasts and on the linguistic experience of the listeners. Wang, Jongman, and Sereno (2001) suggest that since English listeners are accustomed to processing pitch, specifically intonation, in the right hemisphere but lexical stress in the left-hemisphere then both hemispheres are engaged when listening to lexical tones.

Since the domain of English prosodic categories is a phrase, studying English listeners' perception and production of lexical tones within phrase-length utterances may be subject to different interference patterns than isolated syllables. Chen (1997, 2001), White (1981), and Chiang (1979) all suggest that interference or transfer from English intonation can result in errors when learners produce Mandarin tones in longer utterances, since it is only under these circumstances that the native English prosodic system is engaged. Broselow, Hurtig, and Ringer (1987) suggest that interference from the English intonation system can explain their finding that the listeners were less sensitive to the falling tone in non-final position than in final position since in non-final position the falling tone sounds “odd” to native English listeners but is a common English pattern in the final position. However, other studies have found that the falling tone is the most difficult for native English listeners (Shen, 1989; Wang et al., 1999). Chen (1997) found that the position in which a tone was perceived or produced affected accuracy. Tones were generally identified more accurately in initial position than in final position (cf. Miracle, 1989). Furthermore, some general patterns of tonal production were observed in phrases that seemed to correspond to typical English intonation/rhythm patterns. While

these findings are suggestive, the analysis of the production data was based on subjective impressions and the stimuli in the perception test were formed from a very small set that did not carefully control for lexical or segmental factors. Furthermore, no statistical tests were performed on the outcomes.

1.5. Rationale Behind Experiments

The choice of testing the perception and production of Mandarin lexical tones by native Mandarin and native English speakers is motivated by several factors. Mandarin and English represent languages that are very different in their use of pitch both in terms of their types of lexical prosody and intonation and the linguistic meanings of the pitch contours. The contours in the languages are broadly comparable but not easy to directly map onto one another (as may be the case with languages that are structurally and typologically more closely related). Therefore, these two languages offer an ideal testing ground for the investigation of if and how prosodic categories that differ widely influence perception and production for naïve talker-listeners. These studies can be a building block towards the development of a theory of cross-linguistic speech perception and production that includes the perception and production of prosodic structure. In this theory, predictions about both initial perception and the trajectory of learning of non-native prosodic contrasts should be included and this theory should be able to explain how any pair of native and non-native languages will influence one another during cross-language speech perception.

1.6 Overview

A series of three experiments testing the perception and production of Mandarin tones by native Mandarin and English participants is presented in this dissertation. The two central goals of the experiments were to determine the extent to which English listeners rely on linguistic processing (i.e. assimilate the non-native categories to their native ones) and/or auditory processing when listening to non-native prosodic contours and to begin to characterize the influence of contextual variation on non-native prosody perception and production. The first two experiments tested the sensitivity of the listener groups to Mandarin tones in isolation and in tri-syllabic utterances. In Experiment 1, the listeners discriminated pairs of monosyllabic and tri-syllabic stimuli by making same or different judgments. The tri-syllabic utterances included only one tonal frame in which the tones of the first and third syllables were held constant while the tone of the middle syllable varied. This experiment was primarily designed to test the effect on perception of tones presented in isolation versus in longer utterances. In Experiment 2, listeners discriminated tri-syllabic utterances which included three different tonal frames in order to further explore how different patterns of tonal coarticulation affected perception. The responses from the perception experiments were analyzed using both a sensitivity measure and multidimensional scaling. In Experiment 3, participants imitated monosyllabic and tri-syllabic Mandarin utterances as produced by one native Mandarin talker. These productions were identified by native Mandarin judges and submitted to acoustic analysis in order to determine how closely the participants imitated the Mandarin model. The results of the three experiments are used to make recommendations on how

current models of cross-language speech perception might be expanded to suprasegmental perception and production.

CHAPTER 2

2.1 Introduction

One of the central hypotheses in current models of cross-language speech perception is that listeners interpret non-native segmental categories with reference to their native language system of categories. The correspondence between the two sets of categories determines listeners' sensitivity to non-native contrasts as measured by the listeners' ability to discriminate between two contrasting stimuli. The central issue addressed in the first two experiments in this dissertation is whether this hypothesis set forth in the cross-language segmental perception literature can be extended to prosodic contrasts. That is, can the correspondences between native and non-native prosodic categories explain listeners' sensitivity to non-native prosodic contrasts? Or, will other processing factors, such as psychoacoustic processing, primarily affect listeners' sensitivity?

Previous research has shown variability in listeners' abilities to accurately perceive non-native prosodic contrasts. It is important to determine whether this variation can be accounted for exclusively by psychoacoustic factors or whether native language influence also contributes to the patterns of sensitivity to non-native contrasts. If non-native listeners perceive prosodic stimuli in an auditory mode then differences in sensitivity to the contrasts should be primarily related to the acoustic similarity of the categories and the similarity between the native and non-native prosodic categories should not influence listeners' sensitivity. In contrast, if the listeners are processing the

stimuli in a primarily linguistic mode then the assimilation relationships between the categories in the native and non-native languages should primarily account for the differences in sensitivity across the categories. Contextual variation could potentially change both the acoustic similarity between contours and the assimilation patterns between the two languages. Therefore, it is essential to test both the isolated, canonical forms of the prosodic categories and their coarticulated forms. Furthermore, understanding which cues non-native listeners attend to can help to determine how experience with one prosodic system affects patterns of perceptual attention for a new system. Experiments of non-native segmental perception have suggested that listeners attend to cues in non-native segments that are important in distinguishing contrasts in the native language (e.g. Iverson et al., 2003). These cues are not always the ones which most effectively signal the differences between non-native categories. Therefore, differences in attention between native and non-native listeners can lead to differences in identification accuracy or sensitivity. Furthermore, it is logically possible that while non-native listeners' sensitivity to non-native prosodic contrasts remains high, the non-native listeners are attending to different cues than native listeners to make discrimination judgments. In order to gain a full understanding of the ways in which listeners are perceptually processing non-native prosodic contrasts, it is necessary to investigate both sensitivity and perceptual attention. Moreover, investigating attention to acoustic cues may help explain decreased sensitivity to certain prosodic contrasts and can establish the contribution of native language experience to perceptual attention patterns. The use of a discrimination task, in comparison to the use of an identification task, is important for

distinguishing between decreases in perceptual sensitivity and difficulty in placing utterances into abstract categories.

To investigate these issues, the discrimination of Mandarin Chinese lexical tones was tested for native English listeners who had no prior experience with Mandarin or any other tone language. Both native English and native Mandarin listeners were presented with pairs of Mandarin utterances and asked to discriminate between the members of the pairs. These Mandarin utterances included a variety of talkers producing the four Mandarin tones in isolation and in a three syllable utterance in which the first and third tone were held constant and the middle tone varied. Accuracy and reaction time were collected. The accuracy results were used to determine the listeners' sensitivity to the pairs of Mandarin tones. The reaction times were entered into a multidimensional scaling analysis. The similarity spaces for the Mandarin tones in isolation and in the three syllable context for both native Mandarin and English listeners were then determined. From these similarity spaces, the features that native English and Mandarin listeners were attending to when discriminating one and three syllable Mandarin utterances were identified.

The two central goals of this experiment were first to determine the extent to which listeners rely on acoustic versus linguistic processing when listening to non-native prosodic contours and second to examine which acoustic cues were salient to the native and nonnative listeners. If naïve non-native listeners impose the structure of their native prosodic system onto a non-native system of prosodic contrasts then the similarity between native and non-native categories should primarily influence differences in

sensitivity. Specifically, if English listeners impose the structure of the English prosodic system onto the Mandarin lexical tone system then listeners will be less sensitive to contrasts that represent within-category variation in English than to contrasts that map onto across-category variation. Within the Perceptual Assimilation Model, these two types of contrasts would represent single-category assimilation versus two-category assimilation (Best, 1995). Furthermore, if listeners are processing the stimuli in a linguistic mode, then the acoustic cues they attend to should be important cues for distinguishing prosodic categories in their native language. In contrast if listeners are primarily processing the stimuli in an auditory mode then the acoustic similarity among the contours should primarily influence sensitivity and the cues attended to should be psychoacoustically salient and need not be important in distinguishing categories in the native language. The length of the stimuli may also influence what types of knowledge are recruited for discriminating the contours. English listeners may process the longer stimuli in a linguistic mode since the prosodic contours in their language are generally placed over segmental material longer than a syllable but may use primarily auditory processing when listening to isolated monosyllables.

2.2 Method

2.2.1 Stimuli

The stimuli were all consonant-vowel (CV) syllables with the consonant /r/ and the vowel /a/. Both monosyllabic and tri-syllabic stimuli were included. The monosyllabic stimuli included four syllables with the four Mandarin tones. In the tri-

syllabic stimuli, the first syllable was always the high-level tone and the third syllable was the falling tone. The second syllable varied among all four Mandarin tones. A vowel with a high F1 value was chosen so that F0 and F1 were maximally separated facilitating pitch measurements of the stimuli. Additionally, the specific consonant, /r/, and vowel, /a/, were chosen because they do not differ much in their realization between Mandarin and English. Finally, the syllable, *ra,* was deemed appropriate for this experiment because it is not a meaningful word in Mandarin.

The stimuli were produced by three male and three female native Mandarin talkers (average age = 26, stdev = 2.5). The recordings were made on an Ariel Proport A/D sound board with a Shure SM81 microphone in a sound-attenuated booth. After the recording, sound files were converted to the WAV format with a 22,050 Hz sampling rate and transferred to a PC-based computer. The digital speech files were then segmented into individual stimulus files and the root-mean-squared amplitude was equated across all files.

The graphs in Figure 2.1 show averaged curves for the four Mandarin tones in isolation (on the top) and in tri-syllabic utterances (on the bottom). Non-normalized curves for individual talkers are shown in Appendix B. For the purposes of averaging the contours for the figures below, the curves have been normalized for both frequency (f0) and duration. Duration was normalized by either shortening or lengthening the contours in order to have all contours equal to the average duration across talkers for each tone. For the one syllable stimuli, the average syllable duration was taken for each tone and for the three syllable stimuli, each syllable within the utterance was either lengthened or

shortened to equal the average for that syllable. The boundaries between the syllables were determined by observing the waveform, the amplitude contour and the pitch contour and listening to the recordings. The syllable boundaries were typically marked by dips in the amplitude contour and the waveform. Syllables were lengthened or shortened using the enhance-pitch-synchronous overlap-and-add (PSOLA) function in Praat 4.2. This function can change the duration of speech files without modifying the pitch of the speech. Once the contours were averaged in duration, pitch tracks were extracted automatically using the autocorrelation method also in Praat, taking measurements every 10 milliseconds.

F0 normalization was necessary so that different speakers' values could be averaged (particularly across males and females). The values in hertz were normalized for each speaker across the four tones. The following formula was used (Ladd et al., 1985; Rose, 1987; Wang et al., 2003):

$$T=[(lgX-lgL)/lgH-lgL]*5$$

where X is the pitch value in hertz, L is the lowest pitch measurement for that speaker, H is the highest measurement for the speaker. The resulting values range from 0 to 5 and correspond to the pitch values in proposed by Chao (1948) to account for the differences across the four Mandarin tones.

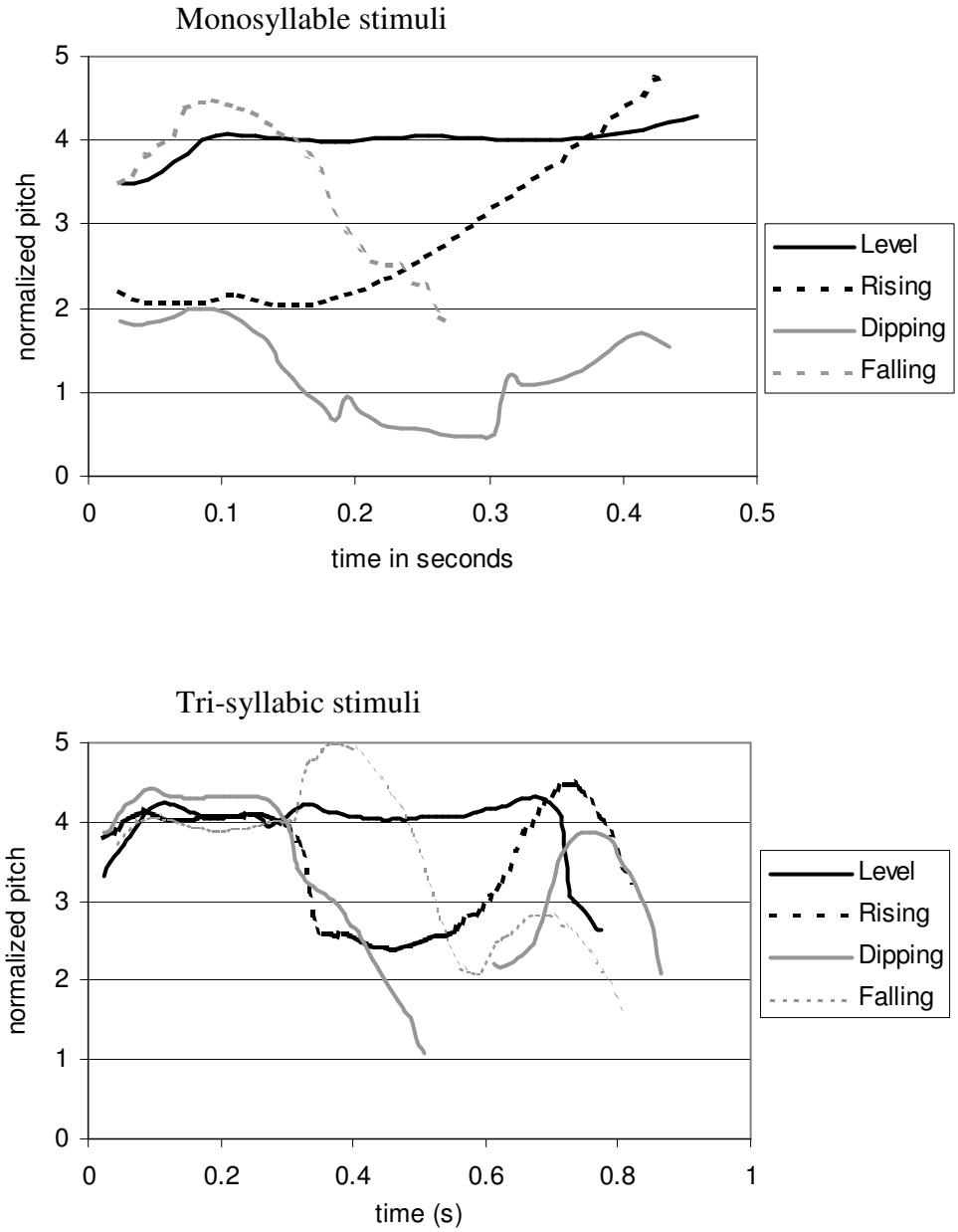


Figure 2.1: Averaged pitch contours for the four Mandarin tones in isolation (top graph) and in the three syllable utterances (bottom graph).

All the stimuli were analyzed in Praat on the following parameters: average pitch, maximum pitch, timing of maximum pitch, minimum pitch, timing of minimum pitch, ending pitch, beginning pitch, pitch range, duration, amount of syllable glottalized (if applicable; both in milliseconds and percent), and timing of amplitude peak. For the tri-syllabic stimuli, all three syllables were analyzed separately. The measurement parameters were chosen based on characteristics found to be important for indicating tone class in surveys of the world's tone languages (Pike, 1948; Ruhlen, 1976; Wang, 1967) and studies of the perception of tonal contrasts (Blicher et al., 1990; Gandour and Harshman, 1978). Some of these acoustic measurements are shown in Table 2.1 below.

Table 2.1: Averaged acoustic measurements for the monosyllabic stimuli and the middle syllable of the tri-syllabic stimuli. Pitch values are given in normalized T values. The timing of the maximum and minimum pitch (% max. pitch and % min. pitch, respectively) are given in percent into the syllable. The percent of glottalization (% glott.) represents the percentage of the syllable produced with glottalized phonation.

number of syllables	Tone	avg. pitch	max. pitch	% max. pitch	min. pitch	% min. pitch	start pitch	end pitch	% glott.
monosyllabic	level	3.99	4.29	66	3.48	29	3.48	4.29	0
monosyllabic	rising	2.84	4.74	92	2.03	24	2.21	4.74	0
monosyllabic	dipping	1.25	2.00	39	0.48	52	1.84	1.53	23
monosyllabic	falling	3.44	4.47	27	1.75	73	3.49	1.75	14
tri-syllabic	level	3.82	4.24	31	3.64	37	3.83	4.24	0
tri-syllabic	rising	2.59	4.01	77	1.68	31	3.58	4.00	0
tri-syllabic	dipping	2.17	3.61	20	1.01	42	2.91	3.61	45
tri-syllabic	falling	3.56	4.60	26	1.88	85	3.72	1.88	3

2.2.2 Participants

The participants were 20 native English (7 male and 13 female; average age = 24.0; age range: 18 – 50) and 20 native Mandarin listeners (8 male and 12 female; average age = 29.1; age range: 22 – 46). All participants had no known speech or hearing impairments. The native English listeners had no prior experience with Mandarin or any other tone language. The native Mandarin listeners all came to the United States as adults and spoke English as a second language. All participants were paid for their participation. Participants were recruited by word of mouth and flyers posted on the Northwestern University campus.

2.2.3 Task

The stimuli were presented in pairs. The listeners' task was to listen to one pair of stimuli and indicate whether they thought the stimuli were the same or different. There were ten possible pairings for the one syllable stimuli and ten pairings for the three syllable stimuli based on each stimulus being paired with itself and all other stimuli with matching syllable count. Each of these pairings was presented six times with presentation order for each pair counter-balanced. The stimuli were presented in 6 blocks of 120 trials each for a total of 720 trials (6 talkers x 10 one syllable pairs and 10 three syllable pairs x 6 presentations of each pair = 720 total trials). Each block consisted of two repetitions of all six talkers' productions of all one syllable or three syllable pairs. The blocks of one and three syllable pairs alternated and half the listeners in each language group were presented with the one syllable block first and half were presented

with the three syllable block first. Between the stimuli in a pair, there was a 350 ms interstimulus interval.

Before the beginning of the experimental trials, the listeners were familiarized with the stimuli by hearing each one once. During the familiarization phase the listeners were not asked to make any responses but just to listen. After the familiarization section, the listeners completed 10 practice trials including all talkers and a mixture of one and three syllable test items to familiarize them with the procedure. Listeners were not given any feedback. Responses were entered on a specially designed response box (SuperLab Pro 2.01). They were instructed to work as quickly as possible without sacrificing accuracy. The testing session lasted approximately 45 minutes. The instructions given to the listeners are shown in Appendix C.

2.2.4 Analysis

The accuracy of the listeners' responses as well as their reaction times were analyzed. Responses were converted to d' scores to assess the listeners' sensitivity to the lexical tone pairs. The d' scores were entered into a repeated measures ANOVA and the appropriate post-hoc tests were performed. d' was calculated by converting the percentage of hits and false alarms to Z-scores under the normal distribution and then subtracting the false alarm rate from the hit rate. Hit rates of 100 percent were converted to 99 percent and false alarms at zero percent were converted to one percent since Z-scores under the normal distribution cannot be calculated for values of zero and 100. More detailed information about this analysis is presented in the results section below.

Reaction times were also analyzed in a repeated measures ANOVA. In addition, non-metric multidimensional scaling (MDS) in SPSS was used to model the perceptual space. Only reaction times for correct 'different' responses were entered into the MDS analysis. Response times for incorrect 'different' responses and correct 'same' responses are not used in this analysis. The average response time for each pair of stimuli for each talker was calculated. Then, the inverse of the average response time of the six talkers was entered into the MDS analysis. The inverse of the reaction time was chosen because reaction times values are approximately reciprocal of distance values (Curtis, Paulos, and Rule, 1973; Shepard et al., 1975). That is, the longer it takes a listener to tell two different stimuli apart, the closer the two stimuli are perceptually. A 4 x 4 dissimilarity matrix was then entered for each subject into an ALSCAL analysis. The Mandarin and English listeners were entered into separate analyses and separate analyses were conducted for each listener group for the monosyllables and the tri-syllabic stimuli. The following options were selected in SPSS. For the level of measurement, ratio was selected because the measured variable is reaction time which has a true zero unlike all other levels of measurement (i.e. ordinal and interval). For conditionality, unconditional was selected, rather than matrix, because reaction times can be objectively compared to one another in contrast to subjective judgments of similarity where all participants may not be using the assigned scale in the same way (Young, 1987). For scaling model, individual differences Euclidean distance model was chosen in order to obtain not only a group solution of the similarity space but also information on how an individual subject's data fit with the group solution. The number of dimensions selected was 2. This

dimensionality was the only possible choice because with an individual differences model one dimension is not possible since subject weights would be undefined. With only four points in the perceptual space, more than two dimensions are not possible.

This technique is limited by only having four points which must be accounted for in a two-dimensional space. With so few points, it can be more difficult to confidently interpret the dimensions and there are a very limited number of ways in which the data can be configured in the space. However, this technique is still useful in attempting to uncover the aspects of the stimuli the listeners are attending to and as a way to visually present reaction times. While a discrimination task can reveal listeners' sensitivity to pairs of stimuli, it does not provide information regarding the acoustic properties of the stimuli on which the listeners are focusing. Therefore, while the MDS analysis is limited, a combination of the sensitivity analysis and the MDS analysis will allow the most comprehensive study of the listeners' responses.

2.3 Results

2.3.1 Sensitivity

Both groups of subjects were quite accurate at the task. The Mandarin listeners had overall accuracy scores of 98% for the monosyllabic stimuli and 97% for the trisyllabic stimuli. While worse overall than the Mandarin listeners, English listeners also were very accurate at this test with 92% correct for the monosyllabic stimuli and 90% correct for the trisyllabic stimuli. Both listener groups on both sets of stimuli were highly accurate at judging pairs of stimuli that were the same: 98% for the Mandarin

listeners for each set of stimuli and 97% for the English listeners for each set of stimuli. From these results, it can be concluded that the listeners understood the task and were performing it as asked. The Mandarin listeners were also highly accurate at judging the different pairs with 98% for the monosyllabic stimuli and 96% for the trisyllabic stimuli. The English listeners were slightly worse at accurately judging stimulus pairs as different particularly for the trisyllabic stimuli: 90% correct for the monosyllables and 86% correct for the trisyllables.

Responses were converted to d' scores to assess the sensitivity of the groups of listeners to the six "different" tonal pairs while controlling for response bias. d' was calculated using the following equation:

d' = Φ (Z-score(hits)-Z-score(false alarms))

In this case, hits were correct responses to different trials (i.e. "different" response to a different trial) and false alarms were incorrect responses to same trials (i.e. "different" response to a same trial). The z-scores were converted from the percent correct scores although scores at the 100% and 0% were converted to 99% and 1%, respectively. This conversion was necessary since z-scores cannot be calculated for 100% and 0% scores. The d' score for one stimulus pair included the hit rate for the pair of stimuli (e.g. correct "different" responses for a level-falling pair) and the average false alarm rate for the two tones included in the different pair (e.g. both the different responses to the level-level tone pair and the falling-falling tone pair).

Both groups of listeners were very sensitive to the tone contrasts. The Mandarin listeners had average d' scores across tone pairs of 4.14 for the monosyllabic stimuli

(standard deviation = 0.36) and 4.01 for the tri-syllabic stimuli (standard deviation = 0.48). The English listeners has average d' scores across tone pairs of 3.59 for the monosyllabic stimuli (stdev = 0.85) and 3.46 for the tri-syllabic stimuli (stdev = 0.92). Figure 2.2 shows the average d' scores for the English and Mandarin listeners on the top and bottom, respectively. The scores are also shown in a table in Appendix D.

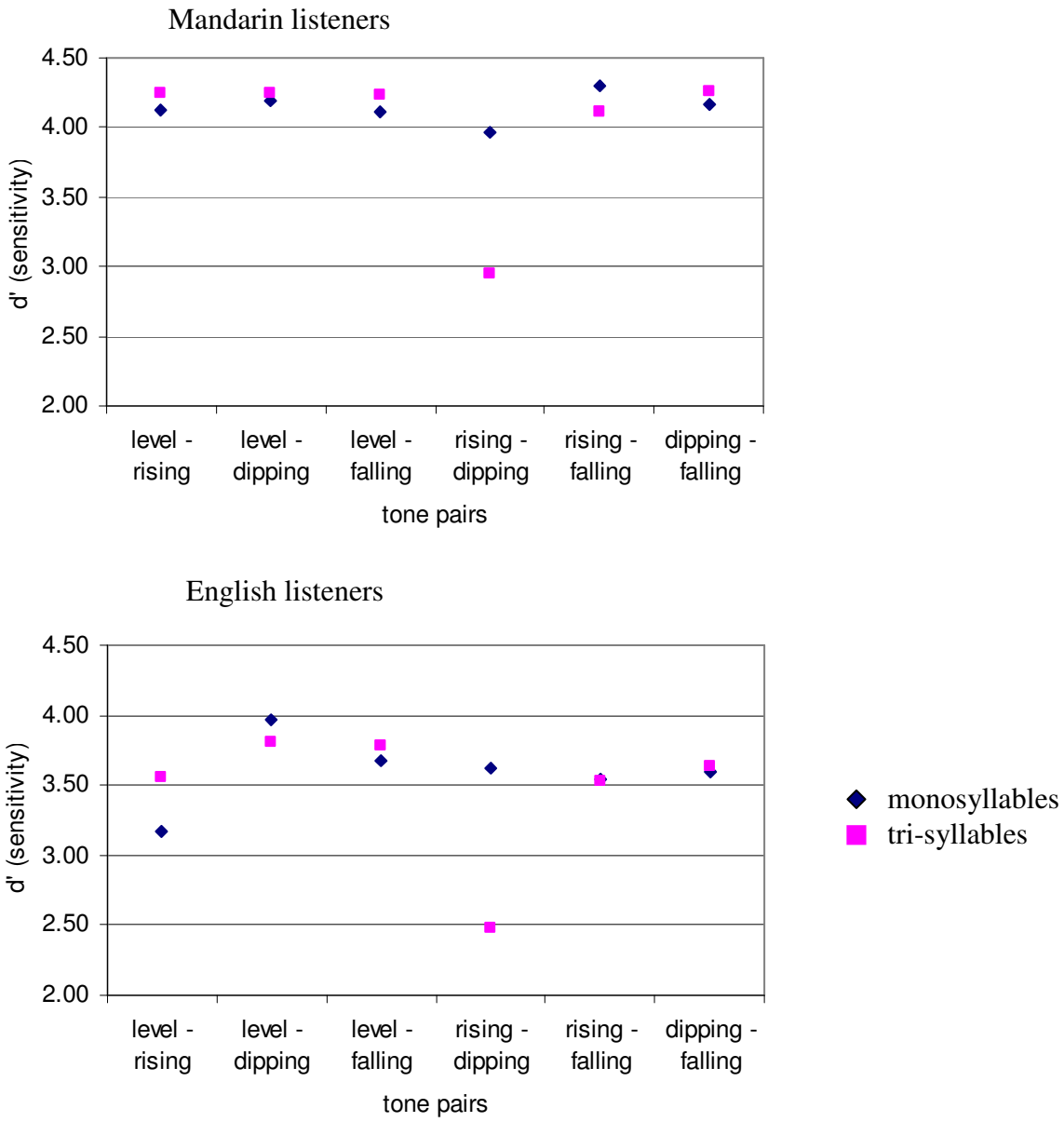


Figure 2.2: d' scores for Mandarin (top graph) and English listeners (bottom graph). Squares show scores for the tri-syllabic stimuli and diamonds show scores for monosyllabic stimuli

The d' scores were analyzed with a repeated measures ANOVA with language (English versus Mandarin) as the between group variable and number of syllables (one versus three) and tonal pairs (level-rising, level-dipping, etc.) as the two within group factors. Significant main effects were observed for language group ($F(1, 380) = 12.49$; $p<0.001$) and tonal pair ($F(5, 380) =56.52$, $p<0.0001$). Furthermore, there were significant interactions of tonal pair and language ($F(5, 380)=5.80$, $p<0.001$) and tonal pair and number of syllables ($F(5, 380)=39.53$, $p<0.0001$). The three way interaction of language, tonal pair and number of syllables was not significant.

In order to determine the cause of the tonal pair and language interaction, the two listener groups d' scores for the tone pairs collapsed across syllable number were compared. The Mandarin listeners were significantly more sensitive than the English listeners to the following tonal contrasts: level-rising ($t(78)=4.05$, $p=0.0002$), rising-falling ($t(78)=3.60$, $p=0.0006$), and dipping-falling ($t(78)=3.86$), $p=0.0002$). The two groups' d' values were not significantly different for any other pairs of tones.

To explore the interaction of tone pair and number of syllables, the d' scores for the tone pairs across syllable number were compared averaging the two listener groups. The sensitivity values for the level-rising pair were significantly higher in the tri-syllabic condition than in the monosyllabic condition ($t(39) = 0.251$, $p= 0.02$). The sensitivity scores for the rising-dipping pair were significantly higher in the monosyllabic condition than in the tri-syllabic condition ($t(39) = 1.079$, $p<0.0001$). All other pairs were not significantly different in the monosyllabic and tri-syllabic conditions.

In sum, while both Mandarin and English listeners displayed high sensitivity to the tone pairs, the Mandarin listeners were more sensitive than English listeners overall. However, the analysis of individual tone pairs revealed that the two listener groups were only different on some of the tone pairs. Collapsing across the listener groups, there were two tone pairs that showed differences across the syllable number conditions. This overall all high sensitivity to the tone contrasts differs from previous studies of English listeners' identification accuracy with Mandarin tones. In Figure 2.3 a comparison of individual Mandarin and English listeners' identification and discrimination scores are shown. The identification scores show a wide distribution for the English listeners which does not overlap with the Mandarin listeners who are essentially at ceiling. The stimuli in the identification task were CV monosyllables with a variety of consonants and vowels produced by one female talker. In contrast, the Mandarin and English listeners' sensitivity scores from the discrimination task show a great degree of overlap.

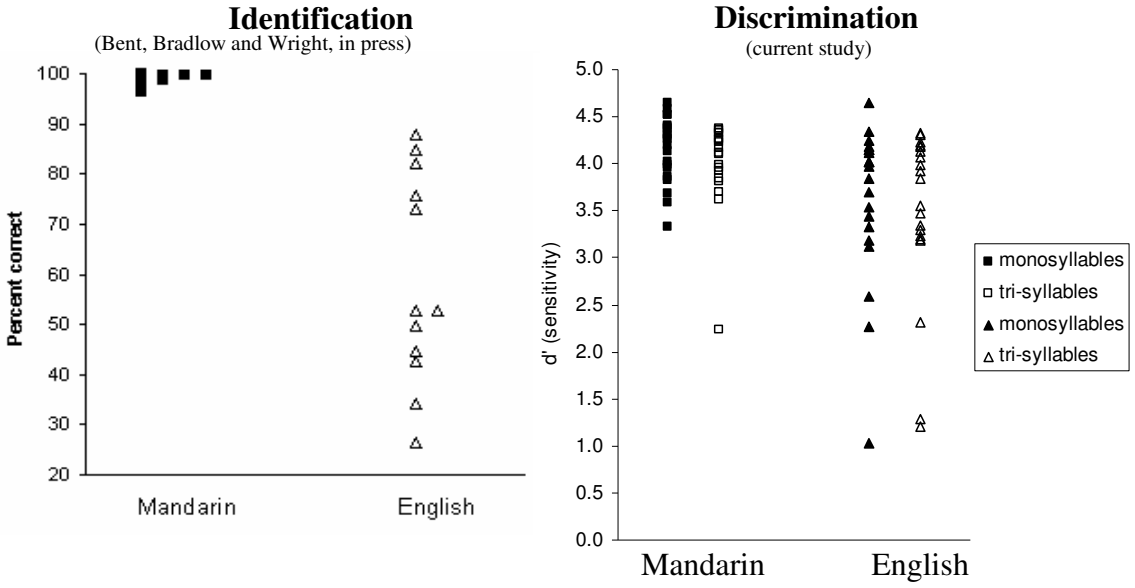


Figure 2.3: A comparison of Mandarin and English listeners' discrimination and identification scores for Mandarin lexical tones. The graph on the left shows average identification scores for monosyllabic stimuli with the y-axis displaying percent correct identifications. The Mandarin listeners are shown in the left column with filled squares and the English listeners are shown in the right column with open triangles. The graph on the right shows the average d' scores on the y-axis across all tone pairs for the monosyllabic and tri-syllabic stimuli in the current study. The squares on the right represent the Mandarin listeners' sensitivity scores and the triangles on the right are for the English listeners. The filled squared and triangles are monosyllabic stimuli and the open squares and triangles are tri-syllables.

2.3.2 Reaction Time Analysis

A repeated-measures ANOVA was performed on the reaction time data (shown in Figure 2.4) with language as the between subjects variable and pair type (level-rising, dipping-falling, etc.) and number of syllables (one or three) as the within subject variables. The only significant main effect was pair type ($F(9, 684) = 14.001$, $p<0.0001$). No other main effects or interactions were significant. Paired t-tests were performed on each of the pair types averaging across the two listener groups and syllable counts. Nearly all "same" pairs had significantly faster reaction times compared to all "different" pairs according to t-tests ($t(79)\geq 3.57$, $p\leq 0.0006$) except that the following tone pairs were not significantly different than the falling-falling pair: level-dipping, level-falling, and rising-dipping. Therefore, while listeners were faster at determining that pairs that were acoustically identical were the same, they did not have difficulty, at least in terms of length of time, with making "different" judgments for any particular tone pair.

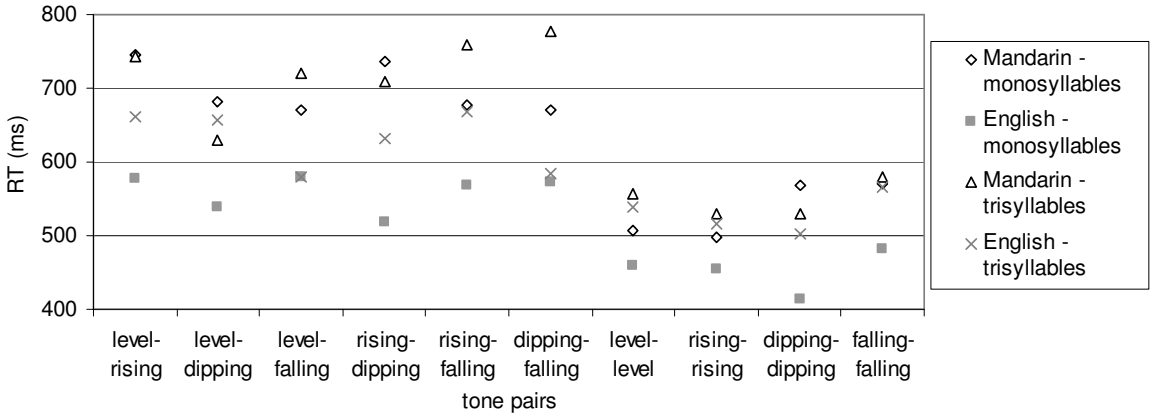


Figure 2.4: Reaction times to tone pairs in each syllable condition by the English and Mandarin listeners.

2.3.3 Multidimensional Scaling

From the multidimensional scaling analyses, four similarity spaces were obtained. These similarity spaces are shown below in Figure 2.5.

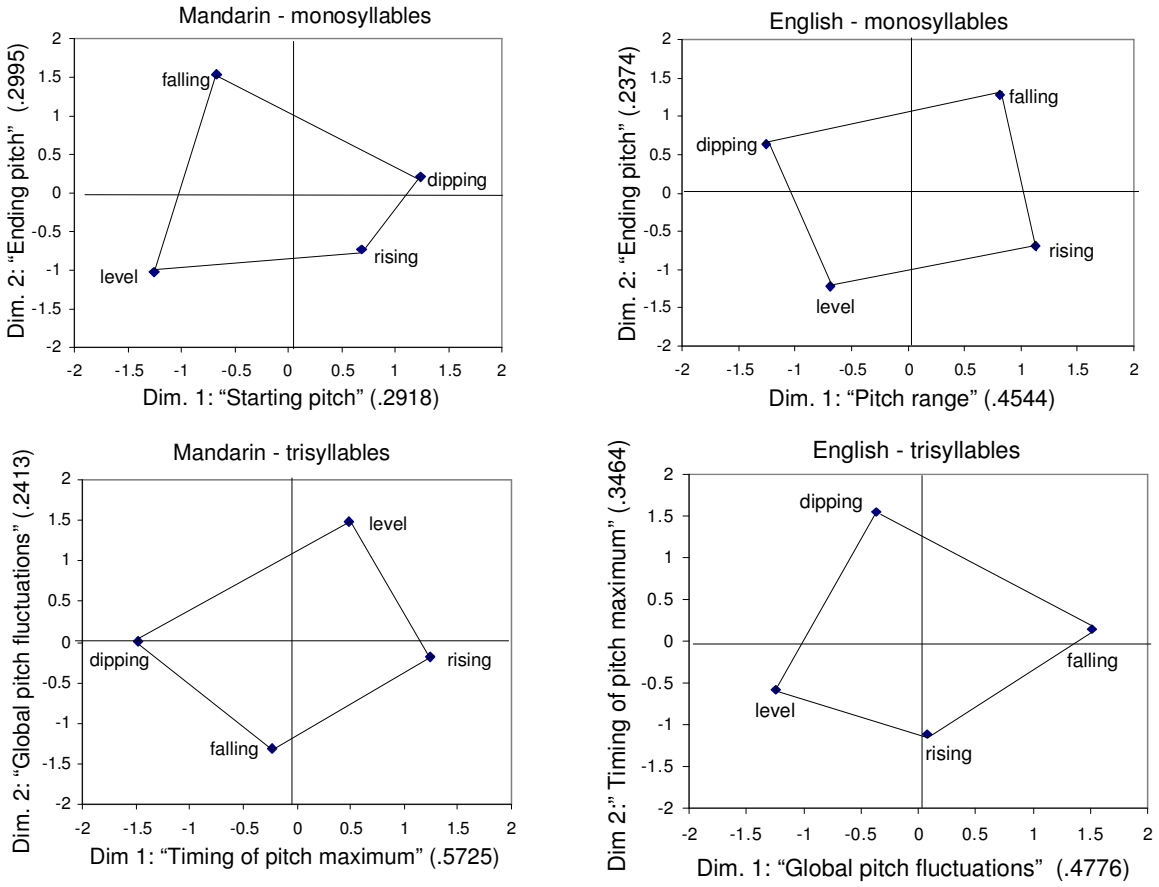


Figure 2.5: Multidimensional scaling solutions for the Mandarin (left column) and English (right column) listeners on the monosyllabic (top row) and tri-syllabic (bottom row) stimuli. Interpretations of the dimensions are shown followed by the averaged subject weights in parentheses (i.e. the average amount of variance accounted for by each dimension).

2.3.3.1 Number of dimensions

With only four points in the similarity space, the maximum number of dimensions that could be determined was 2. A one-dimensional solution was not calculated because subject weights are undefined in a one-dimensional model so the only possibility for the solution was a two-dimensional model. Within this 2-dimensional solution, the R-squared (R^2) value for the four solutions was fairly high (although there was some variance unaccounted for). The R^2 value describes the amount of variance in the scaled distances that the solution accounts for. The R^2 values for the four solutions are as follows: Mandarin monosyllabic condition – 0.591, Mandarin tri-syllabic condition – 0.814, English monosyllabic condition – 0.692, and English tri-syllabic condition – 0.824. These values are a bit lower than the values reported in Huang (2004) for monosyllabic stimuli. In her study, the R^2 value for the Mandarin listeners was 0.89 and the English listener was 0.91. The lower amount of variance accounted for could be related to the greater variability in the current stimulus set since the current study included six talkers and Huang only used one. Furthermore, there were twice as many listeners in the current study as in Huang's.

2.3.3.2 Configuration

The configurations of the monosyllabic stimuli for the Mandarin and English listeners differed in several ways. First, while the tones were not equally separated in perceptual space for the Mandarin listeners (the rising and dipping tones were closer than

any other pair), the distances between the tone pairs for the English listeners were more similar. Furthermore, the Mandarin and English listeners separated the tones in the same way on the second dimension (dipping and falling tones separated from level and rising tones) but in different ways on the first dimension. On the first dimension, the Mandarin listeners grouped the falling and level tones together and the rising and dipping tones together whereas the English listeners grouped the falling and rising tones together and dipping and level tones together.

The configurations of the tri-syllabic stimuli were very different from the configurations of the monosyllabic stimuli for both listener groups. However, the configurations of the Mandarin and English listeners with these stimuli were quite similar, only rotated. For the Mandarin listeners' first dimension, the spaces between four tones were more equal rather than separated into two groups. This ordering was the same for the English listeners on their second dimension. The Mandarin listeners' first dimension, corresponding to the ordering of the English listeners' second dimension, placed the rising and dipping tones in very close proximity with the level and falling tones on either side.

Huang (2004) also reported a closer proximity between the rising and dipping tones for Mandarin listeners than English listeners with monosyllabic stimuli. She interpreted this finding as a confluence of phonetic properties which affected both listener groups' perception and a language specific factor for the Mandarin listeners. Specifically, the tone sandhi rule in Mandarin which links the rising and dipping tones in a phonological alternation pattern was cited as causing the rising and dipping tones to be

perceived as more similar. It is interesting to note that the rising and dipping tones are well separated in the tri-syllabic context suggesting that in a context in which tone sandhi does not apply, these tones may be more perceptually distinct. While the d' score for this tone pair is lower than other pairs, only the reaction times to correctly discriminated trials are entered into the MDS analysis. Therefore, for pairs in which the listeners were able to discriminate the tones, they seem to be well-separated in perceptual space.

2.3.3.3 Interpretation of Dimensions

Using the acoustic analysis of the tones to interpret the dimensions, the ranking of the stimulus points on each dimension was related to a probable physical characteristic of the stimulus. This part of the interpretation is speculative in nature. Since the similarity spaces are based on averaged data across the six talkers, the averaged pitch contours was used to relate the points in the similarity space to the acoustic parameters.

For the Mandarin listeners' solution for the monosyllabic stimuli, the first dimension was found to best correspond to starting pitch. This dimension placed the level and falling tones on one side of the similarity space and dipping and rising tones on the other. The average normalized pitch values for the level and falling tones, 3.48 and 3.49 respectively, were much higher than the values for the rising and dipping tones, 2.21 and 1.84 respectively. The second dimension in this solution was found to best correspond to ending pitch with the falling and dipping tones separated in perceptual space from the rising and level tones. The averaged normalized ending pitch values for

the level and rising tones, 4.29 and 4.47 respectively, were much higher than the ending pitch values for the dipping and falling tones, 1.53 and 1.75 respectively. While this interpretation does not match up exactly with the arrangement in perceptual space, it might be more useful to think of these values as high and low tonal targets with the level and rising tones having high ending tonal targets and the dipping and falling tones having low ending tonal targets. With this interpretation, the difference in the tonal targets separated the rising and level tones from the dipping and falling tones in perceptual space.

For the English listeners' solution for the monosyllabic stimuli, the first dimension was found to best correspond to pitch range with the rising and falling tones on one side of the space and level and dipping tones on the other. The rising and falling tones had much larger pitch ranges, 2.71 and 2.72 respectively than the level and dipping tones, 0.81 and 1.53 respectively. For the second dimension, the best correspondence was ending pitch which was the same as the Mandarin listeners second dimension.

For the Mandarin listeners' solution for the tri-syllabic stimuli, the first dimension was found to best correspond to timing of maximum pitch. On this dimension, the order of the tones was as follows: dipping, falling, level, rising. This ordering in perceptual space corresponded exactly to the ordering of the timing of the maximum pitch in the average contour for the tri-syllabic stimuli. The pitch maximum occurred at the following time points for each of the four stimuli: 94ms (dipping), 380ms (falling), 636ms (level) and 699ms (rising). The second dimension was found to best correspond to global pitch fluctuations. Global pitch fluctuations are measured as the number of pitch reversals in

the pitch contour (i.e. the number of times the direction of pitch movement changed – from rising to falling or vice versa). In the MDS solution, the tones were ranked in the following way: level, dipping/rising, falling. With the slash between dipping and rising indicating that these two were extremely close on this dimension. The tri-syllabic utterance with the level tone had no pitch reversals as it was just flat and then fell. The rising and dipping tones had two pitch reversals and the utterance with the falling tone had three.

For the English listeners' solution for the tri-syllabic stimuli, the first dimension was the same as the second dimension on the Mandarin listeners' solution (global pitch fluctuations) and the second dimension was found to best correspond to timing of maximum pitch and matched well with the Mandarin listeners' first dimension.

2.3.3.4 Weighting of the Dimensions

For the Mandarin listeners in the monosyllabic condition, their first and second dimension received approximately equal weight whereas for the English listeners, their first dimension was much more highly weighted than their second dimension.

For the Mandarin and English listeners in the tri-syllabic condition, their first dimension was weighted more heavily than their second dimension. Interestingly, the dimensions for these two listeners correspond to one another but for the Mandarin listeners, pitch maximum timing gets much more weight than global pitch fluctuations

whereas for English listeners global pitch fluctuations is weighted more heavily than pitch maximum timing.

2.4 Discussion

The current experiment demonstrated both differences and similarities in the ways native English and native Mandarin listeners perceive Mandarin lexical tones in isolation and in tri-syllabic utterances. In the discussion section of this chapter, an explanation of these findings in reference to the structures in the native languages of the two listener groups is provided.

2.4.1 Sensitivity

The Mandarin listeners were overall more sensitive to the Mandarin lexical tone contrasts than the English listeners. However, the English listeners d' values were still quite high demonstrating their great degree of sensitivity to these non-native prosodic contrasts. This high degree of sensitivity may indicate that the listeners are in a psychoacoustic mode of listening in which their linguistic experience does not hinder their ability to discriminate non-native contrasts. This mode of listening corresponds to the nonassimilable category in the Perceptual Assimilation Model. Alternately, if the English listeners perceived the stimuli as linguistic but did not assimilate the stimuli to any particular English prosodic categories then the listeners sensitivity to the stimuli could remain high if the stimuli were acoustically distinct enough. This possibility corresponds to the uncategorizable category in the Perceptual Assimilation Model. The

goal of the following discussion is to determine how psychoacoustic (i.e. language-independent) and linguistic processing interact to produce the reported patterns of sensitivity.

Previous behavioral research with non-tone language listeners' perception of lexical tone has found that while non-tone language listeners are less accurate at discriminating lexical tones, their performance is not at chance. One possible explanation for this result is that listeners are using psychoacoustic knowledge to make their discrimination judgments. For example, Halle et al. (2004) suggest that when non-tone language listeners, in this case French listeners, are identifying or discriminating Taiwanese Mandarin lexical tones they perceive them as "nonlinguistic melodic variations" (p.416) and their perception is psychophysically based. In comparison, Taiwanese Mandarin listeners responses to the tone continua are quasi-categorical. Similarly, Burnham and Mattock (to appear) suggest that tone language listeners base identification and discrimination judgments on linguistic knowledge whereas non-tone language listeners use a more acoustically based strategy. Furthermore, Burnham and Mattock in comparing the relatively poor performance of Australian English speaking adults on the discrimination of a non-native phonemic contrast to the relatively good performance on a non-native lexical tone discrimination task state that "Australian English adults are presumably able to treat the experiment as a perceptual discrimination task with relatively little involvement of linguistic processes, resulting in their relatively better performance on the psychoacoustically salient non-native pitch differences than the non-native consonant difference" (p.5). However, Burnham et al. (1996) found lower

sensitivity to natural speech lexical tone contours than when the same contours were presented either in low-pass filtered conditions or as music suggesting that additional linguistic factors are involved when non-tone language listeners hear lexical tone pitch contours.

Another possible explanation for the English listeners relatively high levels of sensitivity to the Mandarin tone contrasts is that the English listeners are mapping English prosodic categories into tone categories which may help or hinder their discrimination performance depending on the mapping between the two languages. Francis, Ciocca, and Ma (2004) attempted to map Cantonese tones onto English intonation categories to explain the relative ease or difficulty in English listeners' identification and learning of Cantonese tones. Their proposed mappings were based on impressionistic judgments of similarity between the categories. This explanation could help to explain differences in sensitivity among the tonal contrasts as contrasts that represent two-category assimilation would exhibit higher sensitivity than contrasts representing within-category variation or a category goodness difference (Best, 1995; Best et al., 2001). For example, the English listeners have numerically lower d' scores for the level – rising monosyllabic tone contrasts compared to all other monosyllabic tone contrasts. Can the variability in listeners' sensitivity be explained with reference to English intonation categories? One possibility is that English listeners are mapping both of these tones onto the H* pitch accent category which may or may not exhibit a local rise depending on several factors including "degree of prominence of the accent, the distance between the accent and any preceding accent, and the segmental make up of the

syllable" (Ladd and Schepman, 2003, p.82-83). Therefore, for English listeners this pair of Mandarin tones may represent within-category variation for an English intonation category and would present the English listeners with more difficulty in discrimination than other pairs which may represent two-category assimilation patterns.

The other pair of stimuli that presented listeners with the most difficulty was the tri-syllabic rising-dipping pair (i.e. the rising-dipping pair in the level – falling tonal frame). This decrease in sensitivity for this pair of stimuli was observed for both listener groups. Therefore, the explanation for the lower sensitivity to this pair may be the result of psychoacoustic factors rather than linguistic ones. The particular environment in which these tones were placed lead to the syllable with the rising tone starting with a fall (similar to the typical realization of the dipping tone) because it was preceded by the high level tone. Presumably Mandarin listeners are able to compensate for this acoustic similarity in running speech by lexical and syntactic cues but without those cues they are less able to distinguish the tones based on auditory information alone. Presenting the tones in other frame contexts will shed further light on whether this contrast is difficult in context merely because of the particular environment in which it was presented (See Experiment 2 in Chapter 3).

In the cases in which the English listeners displayed the highest sensitivity, the lexical tones seem to map onto two-category variation in English. For example, in the monosyllabic condition, the English listeners' highest d' score was for the level-dipping tone pair. This tone pair may map onto a H* versus L* contrast although neither of the lexical tones are necessarily very good examples of these pitch accent categories. In the

tri-syllabic condition, the level-falling tone pairs displays very high sensitivity (and about equal sensitivity to the level-dipping pair). These two contrasts may also represent across-category variation with the level tone sequence mapping onto something like H* H* L- L% and the falling tone mapping onto a H* L* L- L% sequence.

2.4.2 Multidimensional Scaling

One of the goals of interpreting the multidimensional scaling results is to relate the solution dimensions to features of the listener's language. That is, what about the structure of Mandarin or English would lead listeners to attend to a particular acoustic feature? In both the Mandarin and English monosyllabic solutions, ending pitch was determined to be one of the features that the listeners were using to perceptually separate the tones. In Mandarin, aspects of both the lexical tone system and the intonation system encode important information at the end of utterances/syllables. For lexical tones, underlying pitch targets are best approximated at the end of the syllable (Xu, 2001; Xu and Wang, 2001). Furthermore, Mandarin has "edge tones" within their intonation system that convey information at the end of utterances (Ladd 1996). In English, boundary tones are aligned with the edges of words or phrases. Because every well formed intonation phrase must have a boundary tone, English listeners may have learned through their experience with English to attend to the ends of utterances. Huang (2004) also found that English listeners were attending to ending pitch in the monosyllabic stimuli.

The primary feature in the monosyllabic solutions that the listeners attended to differed between the two groups of listeners. Mandarin listeners attended to starting pitch

whereas English listeners attended to pitch range. For tones in isolation, attending to the starting pitch (in addition to the ending pitch) would provide enough information to separate the four tonal categories as well as provide information about whether the tone was static or dynamic. Huang (2004) also found that Mandarin listeners primarily attended to starting pitch in isolated monosyllables. Pitch range is an important linguistic and paralinguistic feature of English. In terms of the linguistic purposes, Pierrehumbert and Hirschberg (1990) state that "pitch range interacts with the basic meanings of tunes to give their interpretations in context" (p.280) including providing information about the hierarchical structure of the discourse. Furthermore, pitch range provides information about prominence/emphasis and an expansion of pitch range is a cue to narrow focus (Ladd, 1996, p. 202; Ladd and Morton, 1997; Xu and Xu, to appear). The paralinguistic use of pitch range, to express such meanings as emotional state, may be a language general feature (Ladd, 1996).

Since both starting pitch in English and pitch range in Mandarin can convey linguistic information, one remaining question is why the English listeners did not attend to starting pitch and Mandarin listeners did not attend to pitch range. Starting pitch may not convey as much information in English as it does in Mandarin since more components of English intonation align with the end of syllables or phrases than at the beginning. For Mandarin listeners the attention to the pitch range which can convey information about stress or emphasis may not be as salient as information which indicates the lexical tone of a syllable. In Mandarin the lexical meaning of a syllable which

crucially depends on the tone of the syllable may take precedence over the emphasis with which the syllable is produced.

For the tri-syllabic stimuli the two listeners groups attended to the same features of the stimuli: overall pitch fluctuations and timing of pitch maximum. These features are much more global features of the stimuli. That is, listeners need to be attending to the utterance as a whole to determine the realization of the features. For English listeners, the timing of pitch peaks is important for distinguishing pitch accents with high tonal targets. For Mandarin listeners, the location within the utterance of the pitch peak may signal the presence of a high tonal target. As for overall global pitch fluctuations, this feature seems to be related to the similarity between the overall shapes contours. This feature is important for defining overall intonation contours (which map onto differences in pragmatic meaning in English). Mandarin listeners may also be attending to overall pitch fluctuations because these tones are presented in context and Mandarin listeners are also interested in the contribution of the overall pitch contour as it maps onto intonational meaning.

The greater similarity in perceptual attention for English and Mandarin listeners in the tri-syllabic sequences in contrast to the monosyllabic stimuli may be a reflection of the similarities between Mandarin and English in their phrasal intonation categories although they differ greatly in their lexical prosody systems. While all features of the phrasal system can be implemented on a single syllable, these features may become more salient in longer utterances.

2.5 Conclusion

This experiment provides a first step towards the overall goal of extending models of cross-language speech perception to the suprasegmental level. While English listeners were significantly less sensitive than the Mandarin listeners to the lexical tone contrasts, they still demonstrated high sensitivity to all lexical tone contrasts. This result demonstrates that the English listeners experience with English prosody did not result in a complete lack of sensitivity to the Mandarin tone contrasts. Their high sensitivity scores could have either resulted from English listeners ability to use their knowledge of English prosody to distinguish the Mandarin tone contrasts or they could have been operating in a psychoacoustic mode of listening in which they were able to use the acoustic differences among the stimuli to successfully differentiate them.

Because of the overall lack of variability in sensitivity (other than to the tri-syllabic rising-dipping tone pair), it is difficult to relate sensitivity to either acoustic variability or different assimilation patterns between the Mandarin tones and English prosodic categories. Both the Mandarin and English listeners were insensitive to the tri-syllabic rising-dipping tone pair suggests that there may be some tone pairs in which the substantial acoustic similarity leads to lower sensitivity regardless of linguistic experience. Whether the context in which this tone pair was presented caused the lower sensitivity to these stimuli compared with the other tone pairs remains an open question. In Experiment 2 (Chapter 3) other tri-syllabic stimuli will be tested in order to further investigate how tonal coarticulation contributes to variability in sensitivity. The larger number of tri-syllabic stimuli should also allow for more explicit testing of how acoustic

similarity and assimilation to native categories influence listeners sensitivity to non-native prosodic contrasts.

CHAPTER 3

3.1 Introduction

This study, designed to follow-up on findings from Experiment 1 (Chapter 2), investigated how native English listeners' perception of Mandarin lexical tones is affected by tonal coarticulation. Experiment 1 demonstrated that for both native Mandarin and English listeners patterns of sensitivity and perceptual attention to acoustic cues vary depending on whether the lexical tones are presented in isolation or in tri-syllabic utterances. Experiment 1 only used one tonal frame (the level – falling tonal frame) for the tri-syllabic stimuli. Therefore, it is still unknown whether patterns of sensitivity and perceptual attention in longer utterances are common across tones in all tonal frames or if the observed patterns of sensitivity and perceptual attention are specific to the tonal frame tested. In Experiment 2, Mandarin and English listeners' sensitivity and perceptual attention were tested with three different tonal frames.

Mandarin lexical tones vary in their acoustic realizations depending on the tones of the preceding and following syllables. This type of variability is called tonal coarticulation. Tonal coarticulation influenced the results of Experiment 1 as the tones presented in the tri-syllabic utterances were influenced by tonal coarticulation and the tones in isolation were not. The current experiment is designed to more fully investigate how different patterns of tonal coarticulation influence non-native prosody perception. The accuracy with which Mandarin listeners are able to identify a coarticulated tone depends on whether the tone is presented in a compatible or conflicting environment (Xu,

1994). A compatible tone sequence is one in which the end of the preceding tone and the beginning of the following tone match in pitch height. A conflicting tone sequence is one in which the end of the preceding tone and the beginning of the following tone do not match in pitch height. For example, a sequence of a rising tone, a falling tone, and a rising tone is a compatible context whereas three successive rising tones represent a conflicting context. Native Mandarin listeners are more accurate at identifying tones presented in compatible contexts than in conflicting contexts (Xu, 1994).

Very little is known about how surrounding tonal context influences native English listeners' perception of lexical tones. Two experiments (Broselow et al., 1987; Chen, 1997) have tested the perception of Mandarin lexical tones in longer utterances. However, the authors in both of these studies do not report how tonal coarticulation influences perception since the data they report is averaged across different tonal contexts. These studies investigated how a tone's position within an utterance influences identification accuracy. Both Chen (1997) and Broselow et al. (1987) found that a tone's position within an utterance affected English listeners' identification accuracy. Native English listeners made more tone identification errors for Mandarin disyllables in final position than in initial position (Chen, 1997) and were less sensitive to tones in the middle position of a three-syllable utterance than to tones in utterance initial or final position (Broselow et al., 1987). Both studies do not report the listeners' inaccurate responses. While the authors suggest that some error patterns may be the result of interference from English intonation, it is difficult to draw conclusions regarding the relationship between the lexical tones' contour shape and English intonation patterns

since changes in tone realization due to tonal coarticulation were not taken into account. However, these studies demonstrated that a tone's position within an utterance influences non-native listeners' perception. Furthermore, the use of an identification task may have lead to certain confusion patterns that were due to the category names themselves rather than a lack of sensitivity to the pitch contours. For example, native English listeners may inaccurately label the direction of pitch change with English utterances (Arvaniti, personal communication) although they are able to produce them accurately. Therefore, using a discrimination task to test native English listeners' perception of Mandarin lexical tone contrasts will remove the conflation of sensitivity and labeling difficulties. Furthermore, this study will investigate how specific tonal environments influence perception.

To investigate the influence of tonal coarticulation on English listeners' perception, the discrimination of Mandarin Chinese lexical tones by native English listeners who had no prior experience with Mandarin or any other tone language was tested. Both native English and native Mandarin listeners were presented with pairs of Mandarin utterances and asked to discriminate between the members of the pair. These Mandarin utterances included four talkers productions of the four Mandarin tones in three different tonal frames in which the tones of the first and third syllables were held constant while the middle tone varied. Accuracy and reaction time were collected. The accuracy results were used to determine the listeners' sensitivity to the Mandarin tone pairs. The reaction times were entered into a multidimensional scaling analysis. Perceptual similarity spaces were then determined for the Mandarin tones in the three different tonal

frames for both native Mandarin and English listeners. From these similarity spaces, the features that native English and Mandarin listeners attended to when discriminating the Mandarin utterances were identified.

Due to the exploratory nature of this experiment, it was not possible to always make specific predictions. Therefore, while two specific predictions are offered (1 and 2), more speculative possibilities are presented in (3) and (4) below.

(1) As in Experiment 1, English listeners will be overall less sensitive to the lexical tone contrasts than the Mandarin listeners due to their lack of experience with the specific tone categories.

(2) Mandarin listeners should *generally* be able to compensate for the effects of tonal coarticulation and therefore, show relatively consistent sensitivity to tonal contrasts regardless of the frame in which they are presented. However, there may be a few instances in which the changes due to conflicting tonal coarticulation lead to unrecoverability of the underlying tonal targets.

(3) It is not clear how English listeners with no experience with Mandarin tones will handle tonal coarticulation. On the one hand, it is possible that they will show more variable sensitivity to the tonal contrasts because they will be unable to extract the underlying tonal targets from the variable signal. On the other hand, their experience with English phrasal intonation may actually prepare them well for these longer utterances and they may be able to discriminate the patterns of tri-syllabic tones more accurately than has been observed for monosyllabic stimuli.

(4) Mandarin listeners may attend to different acoustic characteristics of the stimuli depending on the environment in which they are presented. Mandarin listeners will have learned through their extensive experience with Mandarin tones which cues reliably reveal tonal identity in particular contexts and therefore will shift attention according to the context. English listeners in contrast will not know which cues reliably distinguish Mandarin lexical tone categories and therefore may attend to the same acoustic cues regardless of the context in which the tones are presented.

3.2 Method

3.2.1 Stimuli

The stimuli were made up of three syllable sequences. All the syllables were CVs with the consonant /r/ and the vowel /a/ resulting in an utterance of the following form: *ra ra ra*. Twelve of these utterances were created using three tonal "frames" (see Figure 3.1). The tonal frames held the tone of the first and third syllables constant while the tone of the second syllable varied among the four Mandarin tones. One frame was a replication condition of Experiment 1 in which the first syllable had the high level tone and the third syllable had the falling tone. In this frame, the level tone was compatible, the dipping tone was conflicting, and the rising and falling tones were half compatible and half conflicting. The other frames included one in which the first and third syllables both had the rising tone and one in which the first syllable had the falling tone and the third syllable had the rising tone. In the rising-rising tonal frame, the falling tone was compatible, the rising tone was conflicting and the level and dipping tones were half

compatible and half conflicting. For the falling-rising tonal frame, the dipping tone was compatible, the level tone was conflicting and the rising and falling tones were half compatible and half conflicting. The motivation for using these three tonal frames was to vary the starting and ending pitches for the middle tone. With the level-falling frame, both the starting and ending pitch targets were high. The rising-rising frame had a high starting pitch and a low ending pitch and the falling-rising frame had low starting and ending pitches. Using these frames resulted in a maximum amount of variability among the target tones in the different tonal frame conditions and would therefore allow a fairly comprehensive characterization of coarticulated tone perception. These utterances were produced by four native Mandarin talkers, two male and two female (average age = 27.3). None of the talkers had participated in Experiment 1. The recording conditions were the same as in Experiment 1.

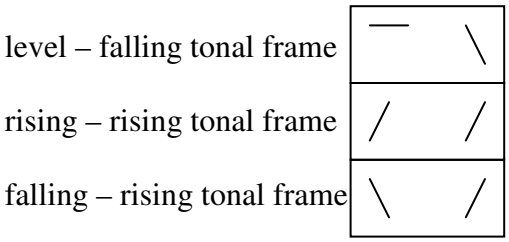


Figure 3.1: Schematic representation of the tonal frames

The three graphs below, in Figure 3.2, show the averaged contours across the four talkers for the three tonal frames. The normalization procedure for the contours was the same as in Experiment 1 as were the types of acoustic features measured. Some of the acoustic measurements are shown in Table 3.1 below. Non-normalized contours for all

four talkers are shown in Appendix E and raw acoustic measurement values for all four talkers are shown in Appendix F.

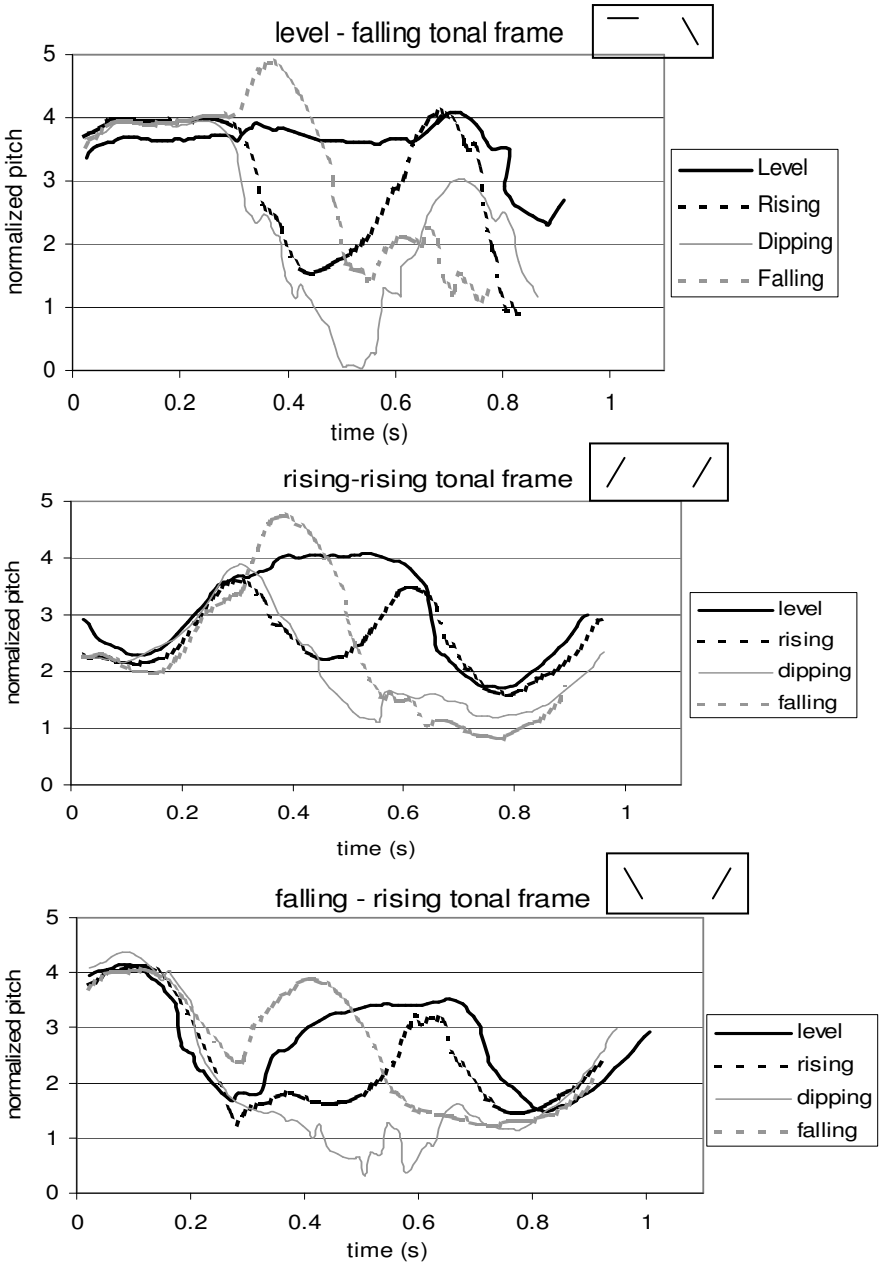


Figure 3.2: Averaged pitch contours for the four Mandarin tones in the level – falling tonal frame (top graph), rising – rising tonal frame (middle graph), and in the falling – rising tonal frame (bottom graph).

Table 3.1: Averaged acoustic measurements for the middle syllable in the three tonal frames. Pitch values are given in normalized T values. The timing of the maximum and minimum pitch (% max. pitch and % min. pitch, respectively) are given in percent into the syllable. The percent of glottalization (% glott.) represents the percentage of the syllable produced with glottalized phonation.

Frame	Tone	avg. pitch	max. pitch	% max. pitch	min. pitch	% min. pitch	start pitch	end pitch	% glott.
level–falling	level	3.7	3.9	15.2	3.5	55.9	3.6	3.6	0.0
level–falling	rising	2.3	3.6	74.3	1.4	37.3	2.8	3.2	0.0
level–falling	dipping	1.4	2.9	4.7	1.5	29.1	2.9	2.1	59.6
level–falling	falling	3.3	4.9	24.8	1.2	91.3	3.9	1.5	0.0
rising–rising	level	3.8	4.2	33.9	2.8	74.6	3.6	2.9	0.0
rising–rising	rising	2.7	3.6	58.5	2.1	33.8	2.9	2.7	0.0
rising–rising	dipping	1.8	3.4	0.8	1.5	39.4	3.2	0.2	43.8
rising–rising	falling	3.0	4.8	21.5	1.1	95.3	3.5	1.1	0.0
falling-rising	level	3.1	3.5	47.4	1.8	0.3	1.8	2.8	0.0
falling–rising	rising	2.2	3.3	73.3	1.3	22.6	1.6	2.7	0.0
falling–rising	dipping	1.0	1.7	35.1	0.2	43.5	1.3	1.2	52.6
falling–rising	falling	2.8	4.0	28.4	1.3	91.7	2.5	1.4	0.0

3.2.2 Participants

The participants were 20 native Mandarin listeners (average age = 26.3; age range = 20 - 33) and 20 native English listeners (average age = 21.0; age range = 18-27). All the requirements for participation and recruitment procedures were the same as in Experiment 1. None of the participants in Experiment 2 had participated in Experiment 1.

3.2.3 Task

The stimuli were presented in pairs. The listeners' task was to listen to one pair of stimuli and indicate whether they thought the stimuli were the same or different. Within each pair, the same talker was always presented. There were 10 possible pairings for each tonal frame based on each utterance being matched with itself and all other utterances with the same frame. Each of these pairs was presented 6 times with presentation order for each pair counter-balanced. The stimuli were presented in 3 blocks of 240 trials each (4 talkers x 10 pairs with the level – falling frame and 10 syllable pairs with the rising – rising frame and 10 pairs with the falling – rising frame x 6 presentations of each pair = 720 total trials). Each block consisted of all the stimuli for one tonal frame. The order of the blocks was counterbalanced within listener groups. Between the stimuli in a pair, there was a 350 millisecond interstimulus interval.

Before the beginning of the experimental trials, the listeners were familiarized with the stimuli by hearing each one once. During this phase, the listeners were not required to make any responses. After the familiarization section, listeners completed 12 practice trials including all talkers and a mixture of different tonal frames to familiarize them with the procedure. Listeners were not given any feedback. Responses were entered on a specially designed response box (SuperLab Pro 2.01). Participants were instructed to work as quickly as possible without sacrificing accuracy. Listeners were given up to three seconds to enter a response. If no response was entered in this time interval, the next trial was automatically presented and "no response" was entered for that trial. The three-second time limit differed from the procedure in Experiment 1 in which

listeners were given unlimited time to respond. In Experiment 1 many listeners had a few outlying reaction time responses. In order to avoid having to determine what length of reaction time counted as an outlier, the response time was limited to 3000 ms in the current experiment. The testing session lasted approximately 45 minutes. The instructions given to the listeners are shown in Appendix G.

3.2.4 Analysis

The analysis procedures were the same as in Experiment 1.

3.3 Results

3.3.1 Sensitivity

Both groups of listeners performed very accurately on the "same" trials with Mandarin listeners responding to 97% of same trials accurately and English listeners responding to 94% of same trials accurately. Mandarin listeners were also highly accurate on the different trials with 97% correct while English listeners were less accurate at 82% correct.

The responses were converted to d' scores to determine the sensitivity of the listener groups to the "different" pairs while controlling for response bias. Both groups displayed high d' values for the three tonal frames. The Mandarin listeners had d' scores of 3.71 for the level-falling tonal frame, 4.11 for the rising-rising tonal frame, and 4.05 for the rising-falling tonal frame when the scores were averaged across tonal pairs. The

English listeners had d' scores of 3.07 for the level-falling tonal frame, 3.11 for the rising-rising tonal frame, and 2.77 for the falling-rising tonal frame when the scores were averaged across tonal pairs. Figure 3.3 shows Mandarin and English listeners' d' scores for the six different pairs in the three tonal frames. d' scores are also shown in Appendix H. d' was calculated in the same way as in Experiment 1.

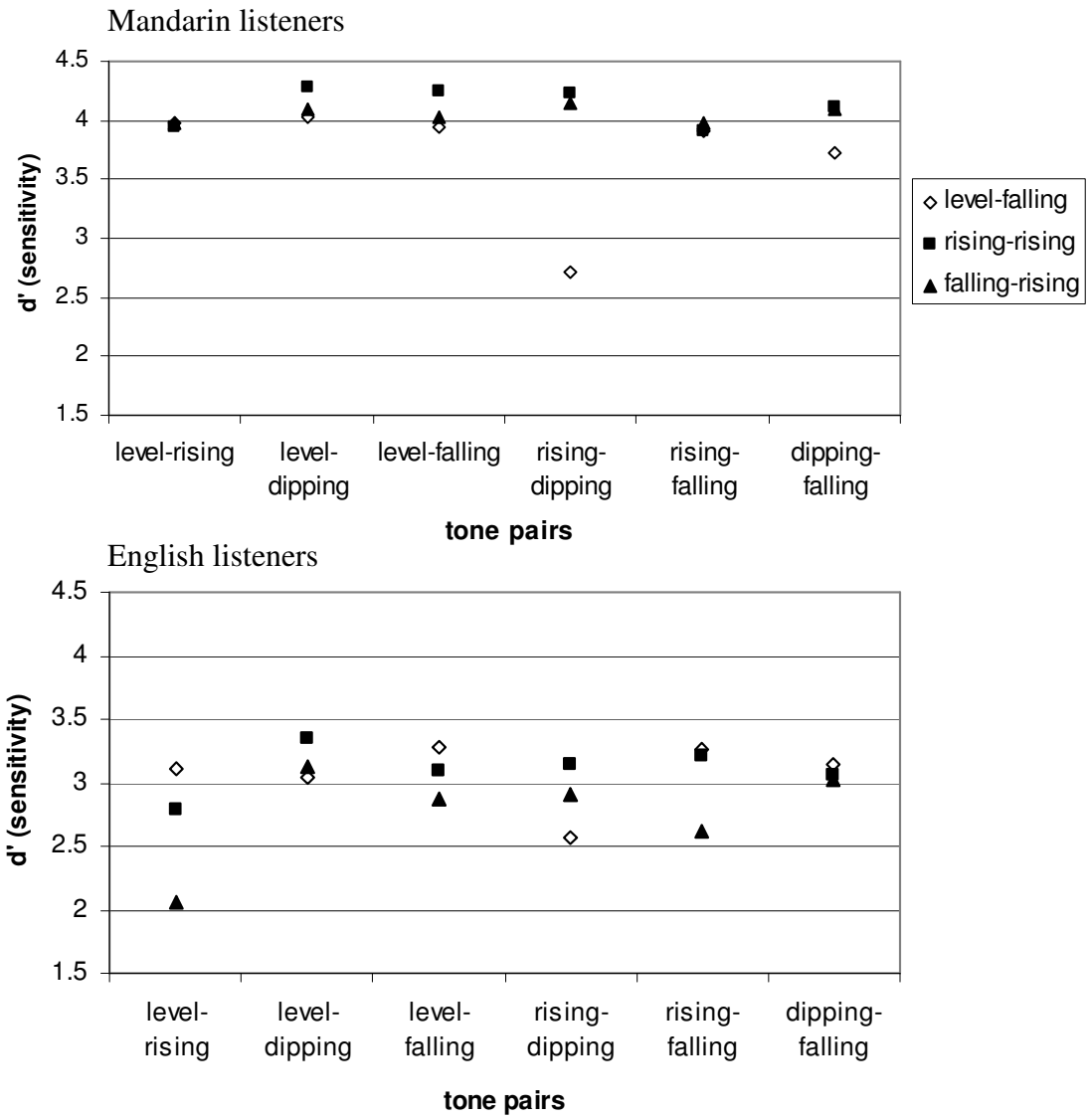


Figure 3.3: d' scores for the Mandarin (top graph) and English listeners (bottom graph). The scores for the tone pairs in the level – falling tonal frame are shown with the open diamonds, the tone pairs in the rising – rising tonal frame are shown with filled squares, and the tone pairs in the falling – rising tonal frame are shown with filled triangles.

The d' scores were analyzed using a repeated measures ANOVA with language (English vs. Mandarin) as the between subjects factor and tonal frame (level-rising, rising-rising, falling-rising) and tone pairs (level-rising, level-dipping, etc.) as the within-subjects factors. There was a significant main effect of language ($F(1, 570) = 42.12$, $p<0.0001$) due to the Mandarin listeners' overall higher d' scores compared with the English listeners. There was also a main effect of tone pair ($F(5, 570) = 16.30$, $p<0.0001$) due to the variability in sensitivity across the tone pairs. The main effect of tonal frame was not significant. The interaction of language and tonal frame was also not significant but all other interactions were significant including the interactions of tone pair and language ($F(5, 570) = 5.52$, $p<0.0001$), tone pair and tonal frame ($F(10, 570) = 17.73$, $p<0.0001$), and a three-way interaction among tone pair, tonal frame, and language ($F(10, 570) = 4.21$, $p<0.0001$).

In order to investigate the cause of the interactions, separate repeated measures ANOVAs were conducted for the two listener groups. For both listener groups, there were no significant effects of tonal frame but there were main effects of tone pair (English: $F(5, 285) = 11.45$, $p <0.0001$; Mandarin: $F(5, 285) = 10.19$, $p<0.0001$) and significant interactions of tonal frame and tone pair (English: $F(10, 285) = 7.81$, $p <0.0001$; Mandarin: $F(10, 285) = 15.23$, $p<0.0001$).

In order to quantify the locus of the interaction between tone pair and tonal frame, for each listener group, the tone pairs were compared across the three different tonal frames. For the Mandarin listeners, the only significant difference across frames was for the rising-dipping pair. The rising-dipping pair had a significantly lower d' score in the

level-falling tonal frame than in the rising-rising ($t(19) = 11.47$, $p<0.0001$) or falling-rising tonal frames ($t(19) = 11.22$, $p<0.0001$). The English listeners' sensitivity to the tone pairs was more affected by the different tonal frames. d' scores for three tone pairs were significantly different across tonal frames. The d' score for the level-rising pair was significantly lower in the falling-rising tonal frame than in the rising-rising ($t(19) = 3.97$, $p = 0.0008$) and the level-falling tonal frames ($t(19) = 4.16$, $p=0.0005$). The rising-dipping pair had a significantly lower d' score in the level-falling tonal frame than in the rising-rising tonal frame ($t(19) = 4.56$, $p=0.0002$). Lastly, the rising-falling pair had a significantly lower d' score in the falling-rising tonal frame than in the level-falling tonal frame ($t(19) = 3.76$, $p=0.001$).

In sum, the English listeners were overall less sensitive to the Mandarin lexical tone pairs than the Mandarin listeners were. Furthermore, while tonal frame only affected sensitivity for one tone pair for the Mandarin listeners, English listeners' d' scores were more widely variable across the tonal frames.

In order to assess the contributions of the acoustic similarity between stimulus pairs to sensitivity, the similarity of the overall shapes of the pitch contours were compared to one another. The relationship between the pairs of lexical tones in the different tonal frames was quantified through a simple correlation between the two pitch contours. These correlations do not take glottalization into account as glottalized values are entered as zero T-values. The differences in duration across the stimuli are taken into account by adding zeros at the ends of the shorter utterances so that all pitch contours have the same number of values. For the English listeners, there was a significant

correlation between the similarity of the pitch contours (as measured by the correlation method just described) and their sensitivity values (Rho = -0.536, $p<0.02$) whereas for the Mandarin listeners there was not a correlation between these two measures (Rho = 0.084, $p = 0.73$) (see Figure 3.4). This negative correlation for the English listeners indicates that as the pitch contours were overall more acoustically similar, the listeners were less sensitive to them.

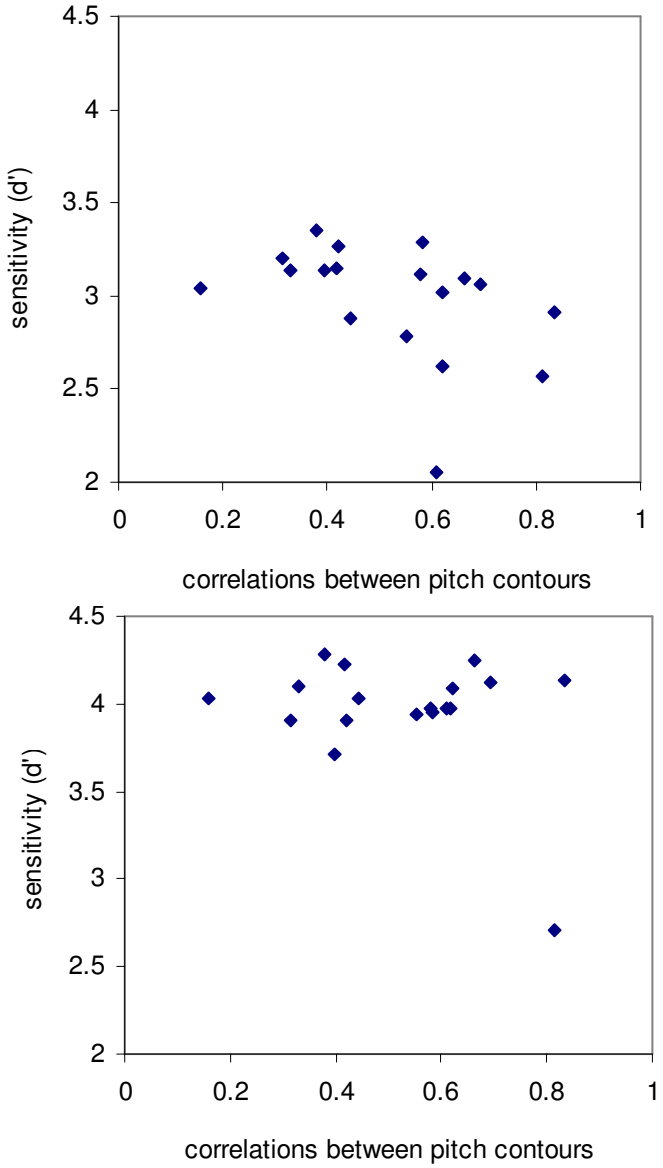


Figure 3.4: Relationship between the similarity of the pitch contours and the sensitivity of the listeners. For the English listeners (shown on top), the two measures are significantly correlated while for the Mandarin listeners (on bottom), these two measures are not correlated.

3.3.2 Reaction Time Analysis

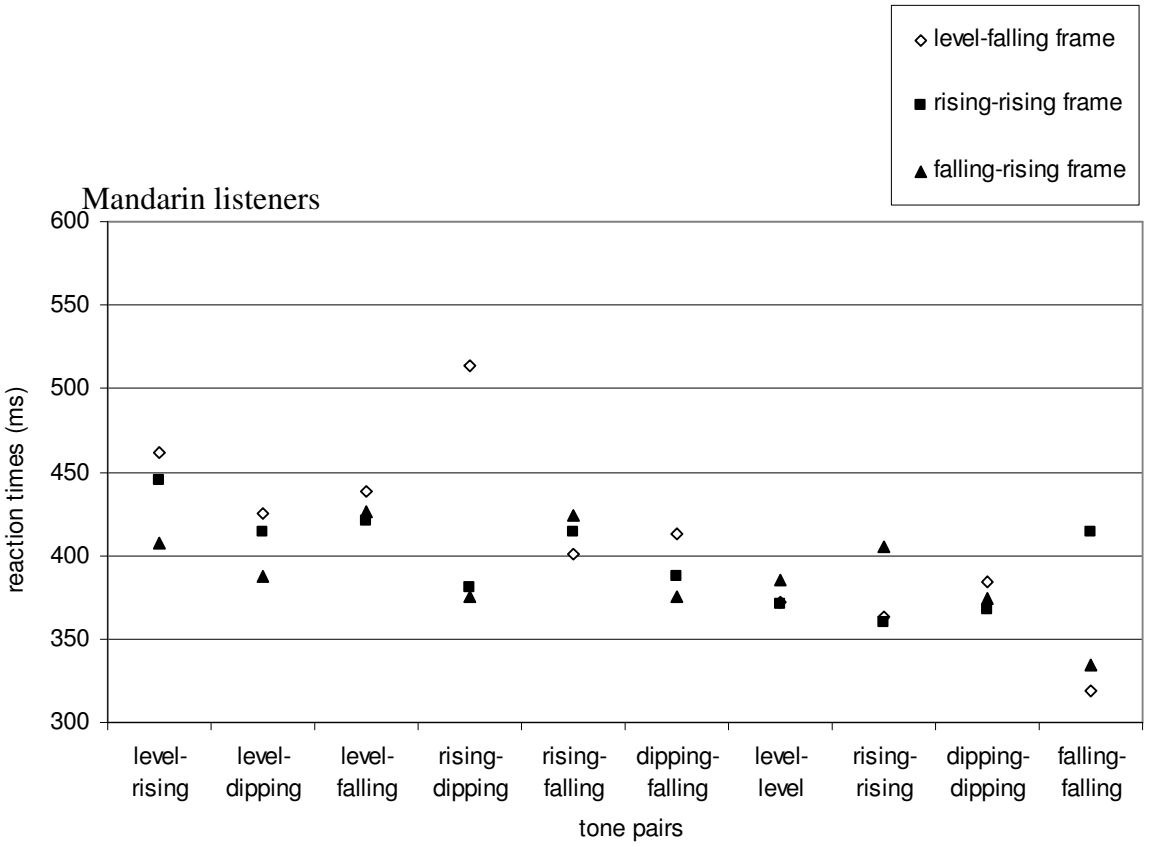
The Mandarin listeners were overall faster at making responses than the English listeners while both groups were faster at responding to "same" trials than "different" trials. The Mandarin listeners' average reaction times were 371 ms for the same trials and 417 ms for the different trials. The English listeners' average reaction times were 439 ms for the same trials and 525 ms for the different trials. The Mandarin and English listeners' reaction times for all tone pairs are shown in Figure 3.5 below.

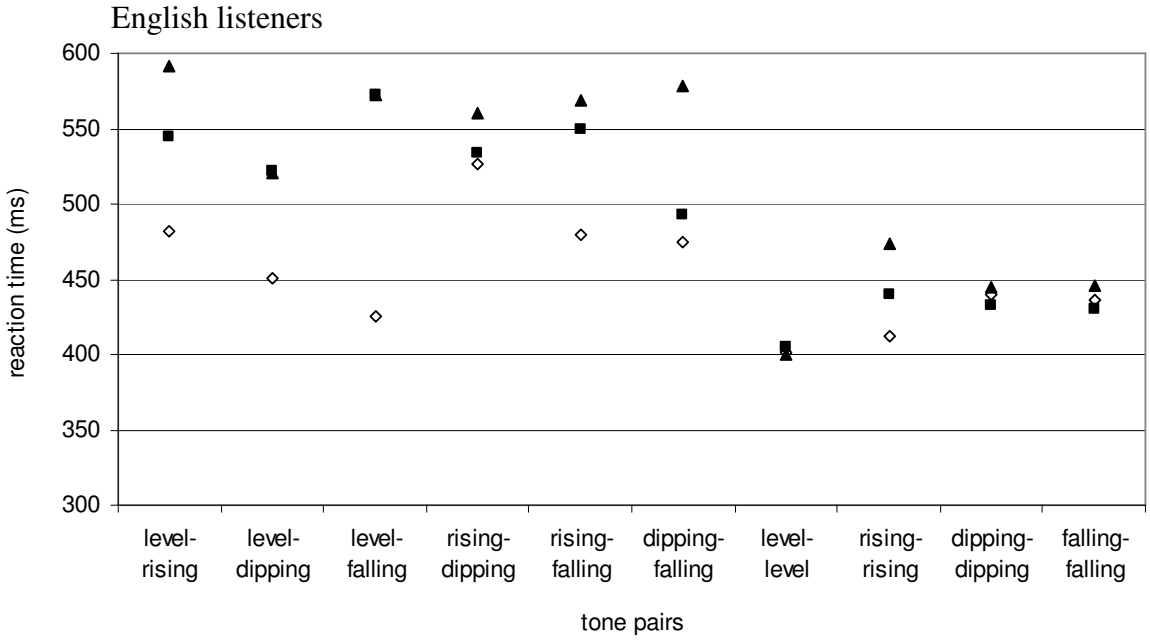
A repeated-measures ANOVA on the reaction time data (to correctly discriminated trials) with language (English versus Mandarin) as the between subjects variable and tone pair (level-rising, level-dipping, etc.) and tonal frame (level – falling, rising – rising, and falling – rising) as the within-subjects variables was performed. Figure 3.4 displays the reaction times for the Mandarin and English listeners for each tone pair in the three tonal frames. There were main effects of both language ($F(1, 1026) = 9.40$, $p = 0.003$) and tone pair ($F(9, 1026) = 13.69$, $p<0.0001$). The main effect of language resulted from the English listeners having overall longer reaction times than the Mandarin listeners. The main effect of tone pair was observed due to the differences among the tone pairs, particularly between the "same" pairs and the "different" pairs. The two-way interaction between tone pair and language ($F(9, 1026) = 2.12$, $p=0.03$) and the three way interaction between tone pair, language, and tonal frame ($F(18, 1026 = 1.71$, $p=0.03$) were also significant.

To investigate the source of the three-way interaction, separate repeated measures ANOVAs for the English and Mandarin listeners were performed. For the English

listeners, there was a significant effect of tone pair ($F(9, 513)=9.93$, $p<0.0001$) but no main effect of tonal frame or an interaction between tonal frame and tone pair. For the Mandarin listeners, there was also a main effect of tone pair ($F(9,513) = 4.61$, $p<0.0001$) and a significant interaction of tone pair and context ($F(18, 513) = 2.33$, $p=0.002$), but no main effect of tonal frame.

Figure 3.5: Mandarin (top graph) and English (bottom graph) listeners' reaction times to the tone pairs in the level – falling tonal frame (open diamonds), rising – rising tonal frame (filled squares), and the falling-rising tonal frame (filled triangles).


level-falling frame
rising-rising frame
falling-rising frame
Mandarin listeners
reaction times (ms)
600
550
500
450
400
350
300
level-rising
level-dipping
level-falling
rising-dipping
rising-falling
dipping-falling
level-level
rising-rising
dipping-dipping
falling-falling
tone pairs


English listeners
reaction time (ms)
600
550
500
450
400
350
300
level-rising
level-dipping
level-falling
rising-dipping
rising-falling
dipping-falling
level-level
rising-rising
dipping-dipping
falling-falling
tone pairs

To further investigate these main effects and the interaction, paired t-tests among tone pairs for the English listener group (averaging across tonal frames) were performed. The only significant differences were between the reaction times of "different" pairs and "same" pairs. While not all the responses to different pairs were slower than responses to same pairs, this pattern accounts for the significant effect of tonal pair observed in the ANOVA. For the Mandarin listeners, the only significant differences across the tone pairs (averaging across tonal frames) were slower response times for the level-rising tone pair than the dipping-dipping pair ($t(59)=3.45$, $p=0.001$) and slower response times for the level-rising, level-dipping, and level-falling were slower than the falling-falling pair ($t(59)\geq 3.47$, $p\leq 0.001$). No reaction times to tone pairs differed significantly across tonal frames.

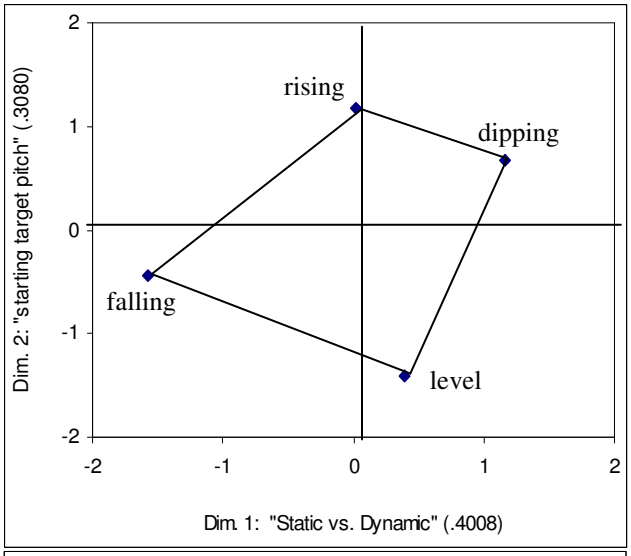
3.3.3 Multidimensional Scaling Analysis

Shown below in Figure 3.6 are the six multidimensional scaling solutions: two solutions, one for each listener group, for each of the three tonal frames. Mandarin listeners' multidimensional scaling solutions are shown in the left column and English listeners' are on the right. The top row has the solutions for the level – falling tonal frame, the middle row has the rising – rising tonal frame, and the bottom row has the falling – rising tonal frame.

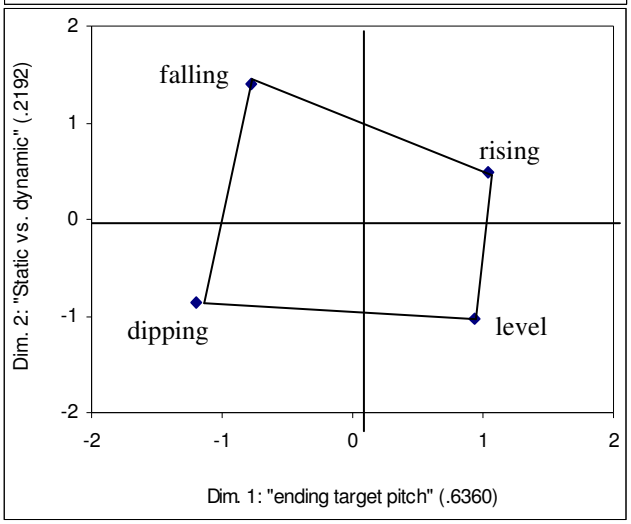
Figure 3.6: Multidimensional scaling solutions for the Mandarin (top three graphs, page 92) and English listeners (bottom three graphs, page 93) for the three tonal frames.

Mandarin listeners

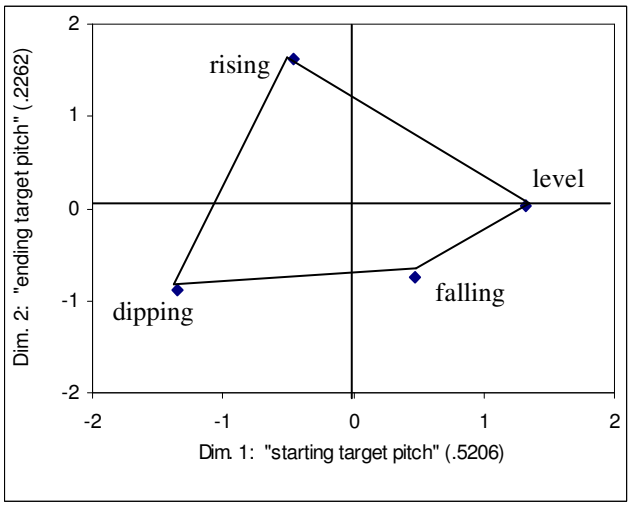
level-falling tonal frame



rising-rising tonal frame

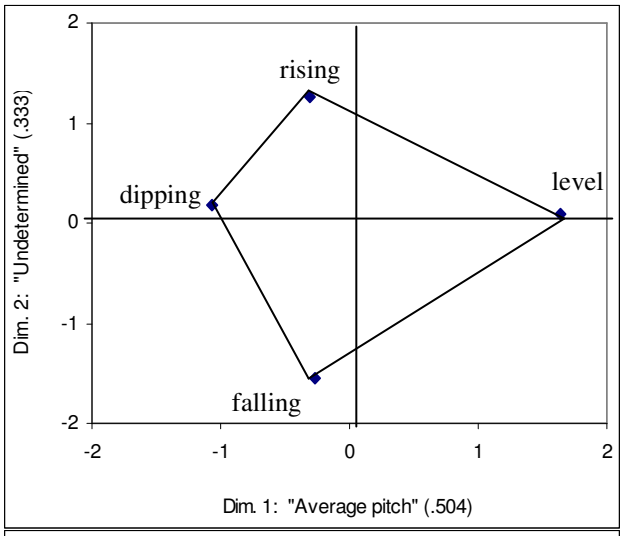


falling-rising tonal frame

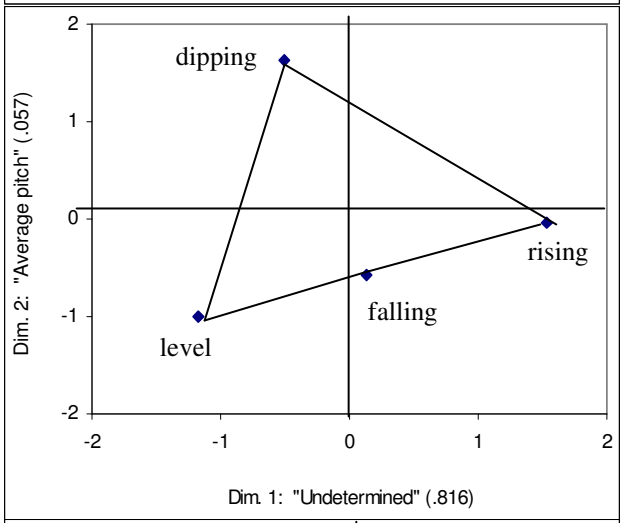


English listeners

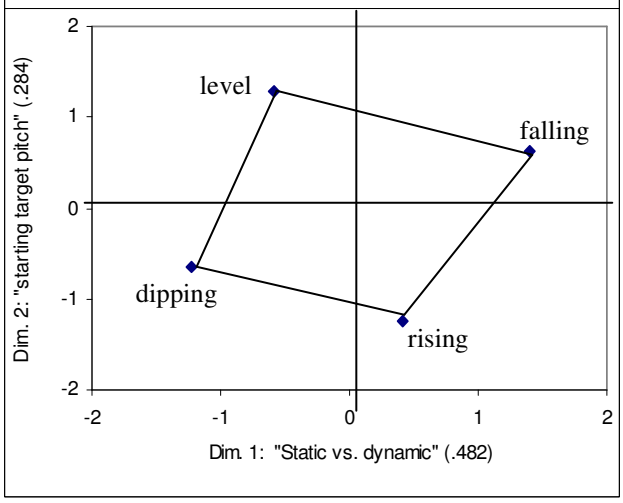
level-falling tonal frame



rising-rising tonal frame



falling-rising tonal frame



3.3.3.1. Number of Dimensions

As in Experiment 1, the only possible dimensionality for the solutions was two. Using an individual differences model does not allow for a one-dimensional solution because the subject weights would be undefined and furthermore, with only four points in each space, a larger number of dimensions is impossible.

3.3.3.2 Configuration

The tones are grouped in different ways on the primary and secondary dimensions across the three tonal frames for both listener groups. For example, while the Mandarin listeners group the tones into rising/level and dipping/falling pairs on the primary dimension for the rising-rising tonal frame, they group them differently, level/falling and dipping/rising, on the primary dimension for the falling-rising tonal frame. Furthermore, the proximity of the tones differs across the tonal frames. In both listener groups' level-falling tonal frame perceptual maps, the rising and dipping tones are closest together in perceptual space but in the other tonal frames, these tones are well separated. These differences in arrangement along the dimensions and proximity of tones in the perceptual space demonstrate that for both listener groups the perceived similarity among the tones varies depending on the tonal frame in which they are presented.

3.3.3.3. Interpretation of the Dimensions

The acoustic features which most closely corresponded to the dimensions in the Mandarin listeners' solutions were tonal targets. All of the dimensions either encoded the

target pitch height at the beginning or the end of the middle tone (high or low) or whether the tone was bi-tonal (a dynamic tone) or had a single tonal target (a static tone). These three parameters accounted for all the dimensions in the Mandarin listeners' similarity spaces. These abstract characteristics seem to capture the arrangement of tones in perceptual space better than any of the acoustic measurements. The static verses dynamic dimension accounts for the primary dimension in the level-falling tonal frame and the secondary dimension in the rising-rising tonal frame. This dimension separates the dipping and level tones (single tonal targets) from the rising and falling tones (bi-tonal targets). The dimension of ending pitch target for the middle tone corresponds to the primary dimension for the rising-rising tonal frame and the secondary dimension of the falling-rising tonal frame. This dimension separates tones ending with the high pitch target (level and rising) from tones ending with a low pitch target (falling and dipping). The starting pitch of the middle tone accounts for the secondary dimension of the level-falling tonal frame and the primary dimension of the falling-rising tonal frame. This dimension separates tones that start with high tonal targets (falling and level) from tones that start with low tonal targets (rising and dipping).

In contrast to the Mandarin listeners, the English listeners do not seem to be attending to tonal targets except in the falling-rising tonal frame. In this frame, the primary dimension for the English listeners corresponds to static versus dynamic tones and the secondary dimension corresponds to middle syllable starting pitch. For the level-falling frame, the primary dimension corresponds to average pitch of the entire contour. On this dimension, the level tone has the highest average pitch (3.57), the rising and

falling tones have similar average pitches (3.01 and 3.16) and the dipping has the lowest average pitch (2.48). The arrangement of the tones on the primary dimension corresponds well to the ranking of the acoustic dimension. For the secondary dimension in the rising – rising tonal frame, the dipping tone is well separated from the other tones that are then ranked from rising, falling, and level. This dimension best corresponds to average pitch of the middle tone with the dipping tone having a lower average pitch (2.1) than the other tones (rising – 2.7; falling – 3.0; level – 3.8). At this point, the secondary dimension for the level-falling tonal frame and the primary dimension for the rising-rising tonal frame remain unresolved as they do not correspond to any measured acoustic features or underlying tonal target pattern.

3.3.3.4 Weighting of the Dimensions

The falling – rising tonal frame was the only frame in which the Mandarin and English listeners were attending to the same cue. Both the English and Mandarin listeners attended to the starting pitch target of the middle tone. However, the Mandarin listeners weighed this cue more heavily than the English listeners did.

3.4 Discussion

3.4.1 Sensitivity

The sensitivity results replicate two of the main findings from Experiment 1. First, the Mandarin listeners were overall more sensitive to the lexical tone contrasts than

the English listeners. Second, while less sensitive than the Mandarin listeners, the English listeners still displayed high sensitivity to the tonal contrasts. These findings reinforce the hypothesis that for English listeners, experience with English prosodic categories does not completely impede their ability to discriminate novel non-native prosodic contrasts. An additional finding from the current experiment was that the English listeners displayed more variable sensitivity to the tone pairs compared to the Mandarin listeners. One of the goals of cross-language speech perception is to determine the cause of the variable ease of perception/acquisition of non-native speech contrasts. This goal is shared by the current investigation. The two most likely causes of the sensitivity variability in the current results stem from auditory or linguistic sources. The differences in sensitivity to tone pairs across the tonal frames may be the result of only acoustic/psychophysical factors. That is, when tone pairs are placed in certain contexts their overall pitch contours become more similar and therefore are more difficult for English listeners to discriminate. On the other hand, these changes in tone realization may lead to differences in the correspondences between Mandarin lexical tone categories and English intonation categories. If, as a result of tonal coarticulation, the pitch contours of a tone pair in a particular context resemble English intonation within-category variation then discrimination should be more difficult than a pair which are assimilated to two English intonation categories. These two potential explanations will be explored below.

A central determinant of the variation in sensitivity across the tone pairs for the English listeners appears to be acoustic similarity since the measures of acoustic

similarity and sensitivity were significantly correlated (See Figure 3.4). However, it should be noted that in previous experiments (e.g. Burnham et al., 1996), English listeners were as accurate as tone language listeners at distinguishing lexical tone-like pitch contours when they were presented as non-speech sounds (i.e. low-pass filtered or music). Therefore, while it appears that the listeners were using primarily acoustic information, the Burnham et al. finding combined with the current one suggest that there is some influence of linguistic processing in the current experiment otherwise the English listeners should display d' values equal to the Mandarin listeners. One possible explanation for the current results is that English listeners process the tones within their native phonological space but they do not assimilate them to any native categories. That is, they are uncategorizable, in PAM terminology, as they are not assimilated into any specific English prosodic category but linguistic processing seems to influence their sensitivity overall. With this type of assimilation pattern, the main determinants of sensitivity are similarity among the pairs and proximity to native categories. Since there is evidence that acoustic similarity, and possibly similarity to native prosodic categories, discussed below, contributed to the observed patterns of sensitivity, it seems that the pairs are most likely uncategorizable for the English listeners but still processed linguistically.

While the correlation between acoustic similarity and sensitivity was significant, there are examples in which the correlation between the two pitch contours are similar and yet sensitivity varies widely. These are cases in which other factors are affecting sensitivity. These factors could include other acoustic properties of the stimuli including amplitude and voice quality differences and/or these differences could be attributable to

different patterns of assimilation or proximity to the native language prosodic categories. Since it seems possible that linguistic knowledge, in addition to psychophysical factors, influences listeners sensitivity, the correspondence of the Mandarin lexical tone categories to English intonation categories was next assessed.

To assess the correspondence of the Mandarin pitch contours to native English intonation patterns, the part of the contour that differed across the tonal pairs was compared to pitch contours for English intonation categories. That is, the contour parts in each tonal frame in which all four tones were realized with the same shape were disregarded in order to simplify the comparisons. The three tone-pairs that had significant differences across the tonal frames were compared to English pitch contours. The rationale behind these comparisons was to determine if Mandarin pitch patterns that were more difficult for English listeners to discriminate would tend to represent pair-wise assimilation patterns with lower predicted sensitivity. Each of the three tone pairs that differed significantly across the tonal frames will be discussed below.

The English listeners' sensitivity to the level-rising tone pair differed significantly across the tonal frames with the falling – rising tonal frame giving rise to significantly lower d' scores than the rising – rising tonal frame or the level – falling tonal frame. In the two frames with the higher d' scores, the two Mandarin contours approximated two different English intonation patterns. In both the level-falling frame and the rising-rising frames, the level tone does not map clearly onto any English contour as it is high and level throughout and the rising tone displays a fall-rise pattern, similar to the L + H* pitch accent in English. This level-rising tone pair in the level-falling and rising-rising

frames may correspond to the uncategorized versus categorized assimilation pattern in the Perceptual Assimilation Model (PAM) in which discrimination is expected to be very good. In contrast, the pattern of the level and rising tones in the falling-rising frame, the one with the lowest sensitivity value, corresponded to the same English intonation pattern: both had a rising pattern or L*+H. This pattern may represent a single-category assimilation pattern in which discrimination is expected to be poor. A comparison of these two tone pairs in two of the tonal frames are shown in Figure 3.7.

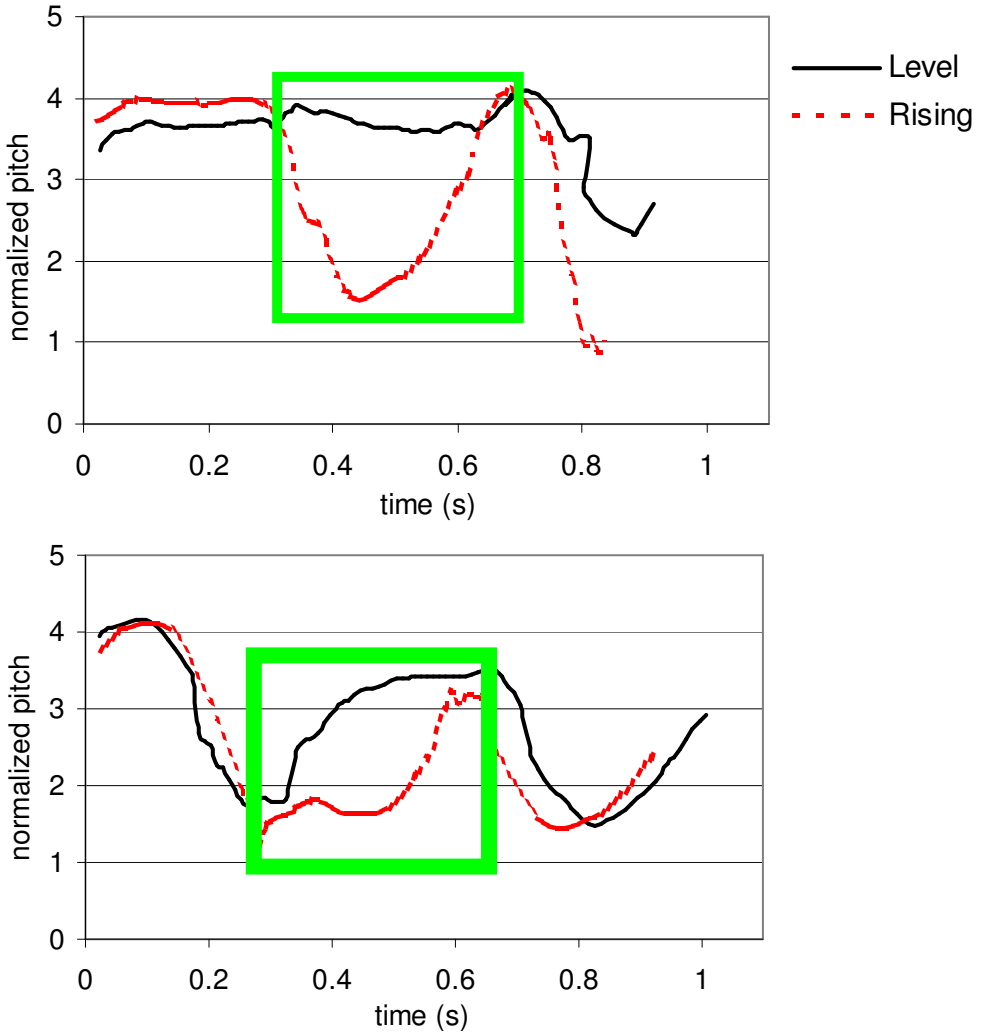


Figure 3.7: The level-rising tone pair shown in two different tonal frames. In the graph on the top, the tone pair is shown in the falling – rising tonal frame in which the English listeners displayed significantly lower sensitivity than in the level –falling tonal frame (shown on the bottom). The pair on the left may correspond to a single category assimilation pattern while the one of the left corresponds to a categorized versus uncategorized pattern.

A similar correspondence between assimilation patterns and sensitivity was observed for the rising-falling tone pair in two of the tonal frames. The English listeners' sensitivity to the rising-falling tone pair was significantly lower in the falling - rising tonal frame than in the level - falling tonal frame. The pattern for both the rising and the falling tones in the falling - rising frame is a rise-fall pattern, similar to the L+H* L* pattern in English. The timing of the peak height varies in the two productions with the falling tone having a peak in the beginning of the syllable and the rising tone having a peak at the end of the syllable but both have peaks within the same syllable. Therefore, this contrast may correspond to a single-category assimilation pattern within the PAM model in which discrimination is predicted to be poor. While discrimination of this contrast would probably not be classified as poor, discrimination of the rising and falling tones in the rising – falling tonal frame was significantly lower than in the level - falling tonal frame. Within the level – falling tonal frame, the rising and falling tones have different pitch patterns. The falling tone has a rise-fall-rise contour (corresponding to the L+H* L+H* pattern) whereas the rising tone has a fall-rise pattern (L* + H). These two patterns correspond to two different English intonation patterns and therefore would map onto a two-category assimilation pattern in which discrimination is expected to be excellent.

The English listeners' sensitivity to the rising-dipping tone pair was lower in the level-falling frame than in the rising-rising frame. In the level-falling frame, the overall pattern for both tones is a fall-rise pattern (L* + H) whereas in the rising-rising frame the patterns are different: fall-rise-fall (L* + H L*) for the rising tone and fall (L*) for the

dipping tone. Again, lower sensitivity is observed for the pair in which the contours map onto a single English intonation category (in the level-falling frame) compared to the pair in which the contours map onto two English intonation categories (in the rising-rising frame). This finding upholds the predictions in the Perceptual Assimilation Model.

This analysis of the mapping between Mandarin and English prosodic categories (lexical tones and pitch accents, respectively) suggests that in addition to the acoustic similarity of the contours, the mapping or proximity between the categories in the two languages may influence sensitivity.

While Mandarin listeners' sensitivity scores did not correlate with the measure of acoustic similarity, the one tone contrast to which they were least sensitive in both the current experiment and Experiment 1, the rising-dipping tone pair in the level-falling tonal frame, is also a pair with one of the highest acoustic similarity scores. The relative low sensitivity for this tone pair seems to be a function of the particular tonal frame in which the tones were presented since Mandarin listeners' sensitivity to the rising-dipping tone pair was high in the other two tonal frames. Furthermore, Mandarin listeners' sensitivity to all the other Mandarin lexical tone contrasts was consistently high regardless of the tonal frame showing an insensitivity to other degrees of acoustic similarity. The rising-dipping pair in the level-falling tonal frame may be a case in which a conflicting tonal context results in unrecoverability of underlying tones due to the very high degree of acoustic similarity. According to Xu (1994), "an underlying phonetic unit is not always fully recoverable when the surface form deviates too much from the canonical form due to coarticulation." In this particular context, the level-falling tonal

frame, the rising tone had a dip in the initial part of the syllable due to carryover coarticulation from the preceding level tone and the dipping tone had a rising portion at the end of the syllable due to anticipatory coarticulation for the high starting pitch height of the following falling tone. These changes resulted in rising and dipping contours that were very similar and may have lead to the Mandarin listeners' inability to recover the underlying tones. Presumably, semantic and lexical information in real world speech communication situations disambiguates these two perceptually confusable tones in this particular context. It would be interesting to investigate real world examples of contexts in which conflicting tonal coarticulation patterns makes tone contrasts ambiguous in order to determine exactly how the contrast is carried acoustically in relation to the semantics and pragmatics of the utterance.

Many of the other tones were also in conflicting contexts but these changes did not significantly neutralize the differences between the tones in a pair. While Mandarin listeners may have more difficulty in *identifying* some tones in conflicting contexts (especially when extracted from the original context) (Xu, 1994), they were very accurate at *discriminating* the tones presented in their original context.

The Mandarin listeners' high sensitivity to the tone contrasts regardless of the acoustic similarity in comparison with English listeners variable sensitivity to the tone pairs depending on their acoustic similarity shows one effect of linguistic experience. That is, the Mandarin listeners are highly sensitivity to nearly all tone pairs because these contrasts represent meaningful contrasts in their native language. In contrast, for the English listeners these contrasts are not meaningful and therefore they may be more

attuned to the gradient similarity among them rather than processing them categorically. For the English listeners these contrasts may either be processed outside of their native phonological space, a nonassimilable pair in PAM terminology, or they may be processed linguistically within the native phonological space but are not assimilated into any native category and therefore uncategorizable. In either of these instances acoustic similarity between the stimuli should correlate with sensitivity. However, since acoustic similarity did not correlate perfectly with sensitivity, there are other factors which influenced the native English listeners' sensitivity. If the categories are within the native phonological space then proximity to native categories can also influence listeners sensitivity according to Best (1995, 2001).

Additional processing demands may influence perception when the meaning of the pitch category, either lexical or pragmatic, is extracted by the listener. Nevertheless, as a starting point, the results from this experiment demonstrate that raw sensitivity to these pitch patterns, devoid of any lexical or pragmatic meaning, is probably influenced by a combination of acoustic similarity and correspondences between prosodic categories in the native and non-native languages. While labeling prosodic categories or extracting meaning from the contours would add additional processing difficulties, listeners' underlying sensitivity is changed as a result of their entrenched native language prosodic system.

3.4.2 Multidimensional Scaling

The Mandarin participants' listening strategy centered on attending to tonal targets. The weight assigned to the particular type of target (starting, ending, or number of targets) varied depending on the context in which the tones were presented. This finding suggests that Mandarin listeners have learned, through their extensive experience with coarticulated tones, which cues will be robust markers of tonal identity in particular environments. The Mandarin listeners are able to recover the underlying tonal targets from coarticulated tones and then they form similarity relationships among the recovered tonal targets.

The dimensions in the multidimensional scaling solutions may be better suited within an autosegmental-metrical (AM) approach rather than a configurational model. In the AM approach all pitch patterns are described as underlying high and low targets. Rising and falling pitch patterns can be explained through the use of bitonal pitch targets. In comparison, the configurational model also postulates high and low targets, but rises and falls are represented directly through underlying rise and fall targets. The similarity between the high tone and the falling tone, for example, can be accounted for as a shared high target in the AM approach but in the configurational approach these two tones would not share any features. Therefore, the interpretation of the dimensions as encoding underlying high or low targets assumes an AM approach rather than a configurational approach. A configurational model would be unable to capture some of the similarity relationships between the tones although it could encode the dynamic/static dimension.

English listeners also appear to be attending to tonal targets but only in the falling – rising tonal frame. English listeners may attend to these targets in the Mandarin stimuli because they also must attend to tonal targets in English (for pitch accents). On the primary dimension, they appear to be attending to static versus dynamic tones corresponding to English single target pitch accents versus bi-tonal pitch accents. The secondary dimension in this tonal frame is starting pitch target (high or low) which also could correspond to high or low targets in English pitch accents. However, in the other two frames, the English listeners are apparently not attending to tonal targets. In both of the other frames, the level – falling tonal frame and the rising – rising tonal frame, the English listeners appear to be attending to average pitch as one of the dimensions. This result conforms to the findings from Gandour and Harshman (1978). They found that English listeners weighed average pitch most heavily during dissimilarity judgments of synthesized speech with a variety of tonal contours. They claimed that this dimension is a nonlinguistic perceptual cue for the non-tone language listeners but could be a linguistic-phonetic cue for the tone language listeners. However, for English listeners, there may be paralinguistic and linguistic uses of average pitch including the encoding of speaker identity and emotion. Therefore, it is difficult to determine whether only psychophysical processing is involved during the attention to average pitch or whether both psychophysical and linguistic processing are activated. Unfortunately, one of the dimensions for both the level – falling and rising – rising tonal frames could not be interpreted. Further research is needed to determine which other acoustic cues the English

listeners are attending to during discrimination judgments of non-native prosodic categories.

3.4.2.1 Comparison of Multidimensional Scaling Results from Experiments 1 and 2

The Mandarin participants' listening strategy of attending to tonal targets found in Experiment 2 for all tonal frames was the same strategy used by the Mandarin listeners for monosyllabic stimuli in Experiment 1. However, the Mandarin listeners used different types of cues for the tri-syllabic stimuli used in Experiment 1 compared to the stimuli in the current experiment. Crucially, in the replication condition in the current study (the level-falling tonal frame), the listeners attending to different dimensions in the two studies. Several differences between the studies may account for these discrepancies but which of these factors lead to the differences is undetermined at this time. One difference between Experiments 1 and 2 was that Mandarin listeners took more time to respond (around 250 milliseconds longer for the stimuli with the level - falling tonal frame) in Experiment 1 than in the current study (Experiment 2). In both studies, the listeners were told to enter a response as quickly as possibly but in Experiment 1 they had had unlimited time to respond while in Experiment 2 they were given a maximum response time of three seconds. These differences in the time course of the Mandarin listeners' responses may have changed their attention to acoustic dimensions. That is, the extra processing time (Experiment 1) may have lead the listeners to attend to different aspects of the stimuli. However, it is possible that other differences across the

experiments may have accounted for the discrepancies in the results including the particular listeners that participated in each experiment and in the implementation of the targets by the specific talkers in each experiment. Furthermore, these differences between the results from the two experiments could indicate the instability of four points in a two-dimensional solution.

The English listeners also attended to different dimensions in Experiments 1 and 2 for the tri-syllabic stimuli in the level – falling tonal frame. Alternately, the differences in the particular listeners or talkers, as mentioned for the Mandarin listeners, may have lead to differences in the results. The main differences between the stimuli in the two experiments are in the pitch range. The stimuli used in Experiment 1 were produced in a slightly compressed pitch range compared to the stimuli in the current experiment. Furthermore, there was less separation between the rising and dipping tones in Experiment 1 than in Experiment 2. Whether these slight differences between the stimuli in the two experiments could have accounted for the differences is unclear.

3.5 Conclusions

The findings from this experiment have demonstrated that contextual changes in the implementation of lexical tones significantly influence non-native listeners' sensitivity to the categories. These changes in sensitivity seem to primarily be the result of differences in acoustic similarity across the tone pairs resulting from tonal coarticulation. In contrast to the results for the non-native listeners, native listeners seem to be relatively insensitive to the acoustic variability which results from tonal

coarticualtion. The significant relationship between acoustic similarity and sensitivity for English listeners suggests that English listeners are primarily processing the tones in an auditory mode with some influence of linguistic processing. The Mandarin listeners, who demonstrated relative insensitivity to acoustic similarity, seem to be processing the tone pairs in a primarily linguistic mode. Evidence for listeners processing non-native speech sounds in a psychoacoustic mode has also been shown for segments. For example, native English listeners appear to perceive Zulu clicks as non-speech sounds (Best, McRoberts, and Sithole, 1988). Furthermore, manipulations of experimental testing conditions in segmental discrimination tasks have shown that non-native listeners utilize different processing strategies depending on the exact experimental conditions (e.g. inter-stimulus interval) (Werker and Logan, 1985). The current results may demonstrate a suprasegmental example in which listeners are recruiting primarily psychoacoustic processing to successfully discriminate non-native contrasts.

The finding that contextual variability significantly influences sensitivity for non-native listeners suggests that models of cross-language prosody perception must incorporate the influence of contextual variation on the perception of non-native prosodic categories in order to fully account for perceptual patterns. The particular ways in which abstract categories are implemented in specific contexts need to be compared across languages in order to fully characterize patterns of cross-language speech perception. While the Speech Learning Model considers context specific allophones when characterizing cross-language comparisons, the Perceptual Assimilation Model and the Native Language Magnet Model both only consider abstract phoneme categories. This

experiment has demonstrated the necessity of incorporating contextual variation into the comparisons across language for suprasegmental categories.

The influence of contextual variation on non-native listeners' perception of prosodic categories is similar to the variation seen in non-native segmental perception. Specifically, non-native listeners' performance differs depending on the placement of the segment within a word (e.g. Logan, Lively, and Pisoni, 1991) and whether the segment is placed in an isolated word or in a sentence (e.g. Strange et al., 2004). While the current study demonstrated that contextual variation influences the perception of non-native lexical tones, further studies should be conducted to assess the contribution of a lexical tone's position within the utterance (building on the work of Broselow et al. (1987) and Chen (1997)). Furthermore, the interaction of segmental and suprasegmental material in non-native speech perception should be investigated.

In this experiment native Mandarin listeners attended to tonal targets in all tonal frames but weighed the specific type of target differently depending on the surround tonal context. This finding suggests that for English listeners to adopt a more native-like listening strategy, they need to attend more to the underlying tonal targets but must be flexible in their attention to specific cues in particular contexts. While training native English listeners to identify monosyllabic Mandarin tones using a high-variability training paradigm has been very successful, future training studies should incorporate longer stretches of speech since English listeners' sensitivity varies depending on the context. Through these training procedures, listeners would need to tune their perceptual

attention strategies according to coarticulatory patterns in order to be successfully identify tones in running speech.

3.6 Future directions

From both the sensitivity data and the multidimensional scaling analysis, it appears that the English listeners may be using some of their experience with English prosody to interpret non-native lexical tone categories. The dimensions suggested by the multidimensional scaling analysis should be studied explicitly through experiments which test the attention to particular acoustic features by pitting various cues against one another with resynthesized speech. Furthermore, studies with listeners from a wider range of languages are needed to tease apart the contributions of stress, intonation, and paralinguistic uses of intonation to the perception of non-native prosody.

Experiments 1 and 2 (Chapters 2 and 3) have tested the discrimination of pitch contours devoid of any lexical or pragmatic meaning. This tact was necessary in order to study naïve listeners' perception of lexical tone contours. However, in addition to differences between the contour shapes and units over which they apply, another substantial differences between the uses of prosody across languages is in the meaning they impart (e.g. lexical or pragmatic). The study of how listeners perceive non-native prosodic categories during a task in which they need to understand the meaning of the category rather than just discriminating the differences between contours is an important step towards fully understanding how two different types of prosodic systems interact during cross-language speech perception.

Testing listeners from other lexical stress languages and other pitch accent or tone languages would help to determine if all listeners would attend to the same features of non-native prosodic categories. If listeners all attended to the same features of non-native prosody, then the observed differences between the Mandarin and English listeners were likely a result of English listeners adopting a language general or psychoacoustic mode of listening. In contrast, if different patterns of sensitivity and perceptual attention were observed among listeners from a variety of languages, then this would more strongly suggest that the structure of the native language prosodic system sculpts the way listeners perceive non-native prosodic contours. For example, listeners from a language without edge tones (such as possibly Yoruba, Ladd (1996)) would be expected to rely less on ending pitch (as was observed in both the Mandarin and English listeners in Experiment 1) whereas listeners from a language like Hmong in which voice quality is used contrastively would presumably rely more heavily on voice quality.

While these experiments provided some tentative evidence that the structure of English intonation influences the perception of non-native lexical tone categories, further studies with a wider variety of languages will be needed in order to determine which perceptual biases arise from language universal factors and which from language specific factors. Testing languages that have a more limited number of intonation contours compared to English, such as Hungarian, might also narrow down the possible interactions between native and non-native prosodic categories. It would be of interest to test listeners from a system without lexical stress (e.g. French, Bengali or Korean) in order to determine how the prosodic features of stress, intonation and lexical tone interact

in cross-language speech perception. That is, since English has both lexical stress and intonation which interact to produce the pitch patterns in the language is it difficult to determine which parts of the prosodic system are influencing their perception of the Mandarin lexical tones. Testing listeners from native languages which systematically vary in their similarity or difference to the dimension of the non-native prosodic system would help to establish which part(s) of their native prosodic system listeners use when encountering a non-native system.

CHAPTER 4

4.1 Introduction

Both general theories of speech perception and experimental evidence on the representation of non-native speech categories posit links between perception and production categories (e.g. Liberman and Mattingly, 1989; Fowler, 1986; Best, 1995; Kuhl and Meltzoff, 1996; Newman, 1996). However, findings suggest that when there is a discrepancy between the two modalities perception leads production (e.g. Flege, 1988). Therefore, the perception and production of non-native prosodic categories should be related, but the accuracy with which the non-native prosodic categories are produced may be limited by perceptual accuracy. Furthermore, studies of non-native segmental categories have found that the categories of the native and non-native language influence one another (e.g. Flege, 1987). If this relationship can be extended to the production of prosodic categories then non-native speakers' deviations should be shifted towards the norms of their native language.

In cross-language research, a wide variety of studies have shown unified representations or positively correlated abilities for perception and production. The tie between perception and production categories in the non-native language had been observed for sentences (Flege, 1988), vowels (Flege, MacKay, and Meador, 1999; Flege, Bohn, and Jang, 1997; Levey, 2004), and consonants (Flege, 1993; Flege and Schmidt, 1995; Schmidt and Flege, 1995; Bradlow et al., 1997). This relationship has been observed for late-acquiring non-natives and early bilinguals from a wide variety of native language backgrounds. The transfer of learning from one modality to the other in non-

native segmental training studies has provided further evidence for unified representation in these two modalities. The learning acquired through non-native consonant identification training has transferred to more native-like production of those consonant categories (Bradlow et al., 1997; Rochet, 1995). While the representations within these modalities are linked, the acquisition of perceptual categories seems to lead accurate acquisition of production categories.

The Speech Learning Model (Flege, 1995) is the only model of cross-language speech perception which makes explicit predictions about production and the perception-production relationship. The model includes the following hypotheses about non-native speech production. First, the SLM predicts that production of a speech sound eventually corresponds to the perceptual representation of the category. Second, the model hypothesizes that when a sound in the second language is perceived to be very similar to a sound in the first language, these sounds may form a merged representation during acquisition of the second language which does not adhere to the native language norms in either language. These merged phonetic categories will resemble one another in production. Lastly, the model explicitly states that learners' abilities to perceive the differences between non-native phonetic categories constrain their production accuracy.

While much less is known about the perception-production relationship within the prosodic domain, two training studies on the acquisition of Mandarin tone for non-natives have observed transfer of learning between perception and production. Wang et al. (2003) trained native English speakers to identify the four Mandarin lexical tone categories and found that the participants' perceptual identification improvements

transferred to more native-like productions. Furthermore, the rank ordering of tone confusions in perception and production were highly correlated (in both the pre-test and post-test phases of their training study). In Leather's (1990) training study in which native Dutch speaking participants were trained on either production or perception of Mandarin lexical tone categories, the training in one domain transferred to the other suggesting shared representations for production and perception. In contrast, Chen (1997) reported no relationship between accuracy in production and perception for native English participants who had been studying Mandarin for approximately one year. This null effect may be due to the impressionistic ratings of production accuracy without accompanying acoustic analysis and the differences between the stimuli used for the perception and production phases of the experiment.

Many more experiments have been conducted on non-native Mandarin tone production without an investigation of the listeners' perception abilities. These experiments have not always been in accord with each other on the quantity and quality of the errors found. These discrepancies between experiments are most likely due to the wide variation across studies in the type of stimulus materials, methodologies used in eliciting the productions, the listeners' experience with Mandarin, and data analysis methods. The stimulus materials used in the studies of non-natives' productions of Mandarin tones have included reading isolated monosyllables (Bluhme and Burr, 1972; Leather, 1990; Wang et al., 2003), reading sentences (Miracle, 1989), reading a paragraph (Shen, 1989), and spontaneous speech on a specified topic (Chen, 1997). All of these studies used real words as stimulus materials. While some of these studies looked

at tone production in longer utterances (Shen, 1989; Chen, 1997; Miracle, 1989), there was no control for the surrounding context as the analyzed words were considered apart from the surrounding tonal context. Furthermore, the ways in which the data were analyzed varied from using informed native speaker judges, naïve native speaker judges, and acoustic analysis. The experience of the participants ranged from none to about one year of classroom instruction in Mandarin.

While the studies cited above reached different conclusions about which of the Mandarin lexical tone categories were the most difficult for native English speakers to produce, several types of errors commonly emerged. The production errors recorded in these studies included the following: (1) pitch range errors – non-native talkers generally produced a pitch range that was smaller than native Mandarin talkers and/or produced less extreme rises or falls (Leather, 1990); (2) tone register errors – non-native talkers' productions did not match with native Mandarin talkers in terms of pitch height particularly for the beginning and ending points of the syllable (Leather, 1990; Miracle, 1989; Wang et al., 2003; Chen, 1997); (3) contour errors – the wrong contour was produced by either producing the incorrect direction of pitch change or by substituting a static tone for a dynamic tone or vice versa (Miracle, 1989); (4) incorrect placement of the turning point for the dynamic tones (Wang et al., 2003). In addition to these general patterns of errors, a tone's position within an utterance also affected its production. Miracle (1989) found that errors were fewer in final position than initial position.

While the hypothesis that native English speakers' production errors are caused by interference from English intonation has been put forth (e.g. White, 1981; Chen,

1997), no systematic comparison of native English speakers' deviant productions of Mandarin tones to English intonation categories within a theoretical framework has been provided. For example, Chen (1997) observed particular sequences of tones that seemed to correspond to English intonation patterns but relied on impressionistic observations to make this assertion.

In the current study, the relationship between the perception and production of Mandarin lexical tones by naïve native English participants was examined. Native English speakers were recorded imitating both monosyllables and tri-syllabic utterances as produced by one native Mandarin speaker. The monosyllables included all four Mandarin lexical tones and the tri-syllables included utterances in which the tones of the first and last syllables were held constant while the tone of the middle syllable varied between the four Mandarin lexical tones. The native English participants' productions were identified by native Mandarin listeners and were submitted to acoustic analysis. These two methods of analysis were used to determine the ways in which the native English participants' productions deviated from the Mandarin model and whether tones that are distinct perceptually are also well separated in production.

Guided by findings from previous cross-language research with segmental and prosodic perception and production and hypotheses put forth in the Speech Learning Model, the following specific predictions were made:

(1) The Mandarin lexical tones that the native English listeners had difficulty discriminating in Experiment 1 will also be difficult for them to produce. For example, the English listeners had the most difficulty discriminating the level and

rising tones in the monosyllabic stimuli and the rising and dipping tones in the tri-syllabic stimuli. Therefore, the participants should be limited in their abilities to accurately distinguish these tone pairs in production.

(2) Mandarin lexical tones that were discriminated well by the native English participants may not be well separated in production. The finding that the participants were very sensitive to certain contrasts, for example the monosyllabic level-dipping contrast or tri-syllabic level- falling contrast, does not mean that these contrasts will be well separated in production.

(3) English listeners lexical tones productions that significantly deviate from the Mandarin targets will be shifted towards English production norms. These norms may include broad differences between the two languages including the differences in pitch range (generally wider in Mandarin than English) or rate of pitch fluctuations (greater in Mandarin than in English). The prosodic categories of the two languages could also interact including the Mandarin lexical tone categories and the English pitch accent categories. For example, while Mandarin has a high level tone, English does not have any such pattern but does have a level pattern in the middle of the pitch range. Therefore, the production of the Mandarin level tone may shift towards a more English-like production in the middle of the pitch range.

4.2 Method

4.2.1 Participants

The participants were the 20 native English participants from Experiment 1 (7 male and 13 female; average age = 24.0; age range: 18 – 50). Participation in this experiment always occurred directly after the completion of the perception task. All participants had no known speech or hearing impairments and had no prior experience with Mandarin or any other tone language. All participants were paid for their participation. Participants were recruited by word of mouth and flyers posted on the Northwestern University campus.

4.2.2 Production Models

The production models were the same as the stimuli used in Experiment 1. That is, the production models were all CV syllables with the consonant /r/ and the vowel /a/. There were four monosyllabic stimuli and four tri-syllabic production models. The monosyllabic production models included four monosyllables with the four Mandarin tones. In the four tri-syllabic production models, the first syllable was always the level tone and the third syllable was the falling tone. The second syllable varied among all four Mandarin tones. One male native Mandarin talker's productions of these materials were used as production models. This talker was one of the talkers used to create the Experiment 1 perception stimuli. This talker was selected due to his productions having the least glottalization, the least fricative-like /r/s, and durations that were average

compared to the other talkers included in Experiment 1. This talker's productions are shown below in Figure 4.1.

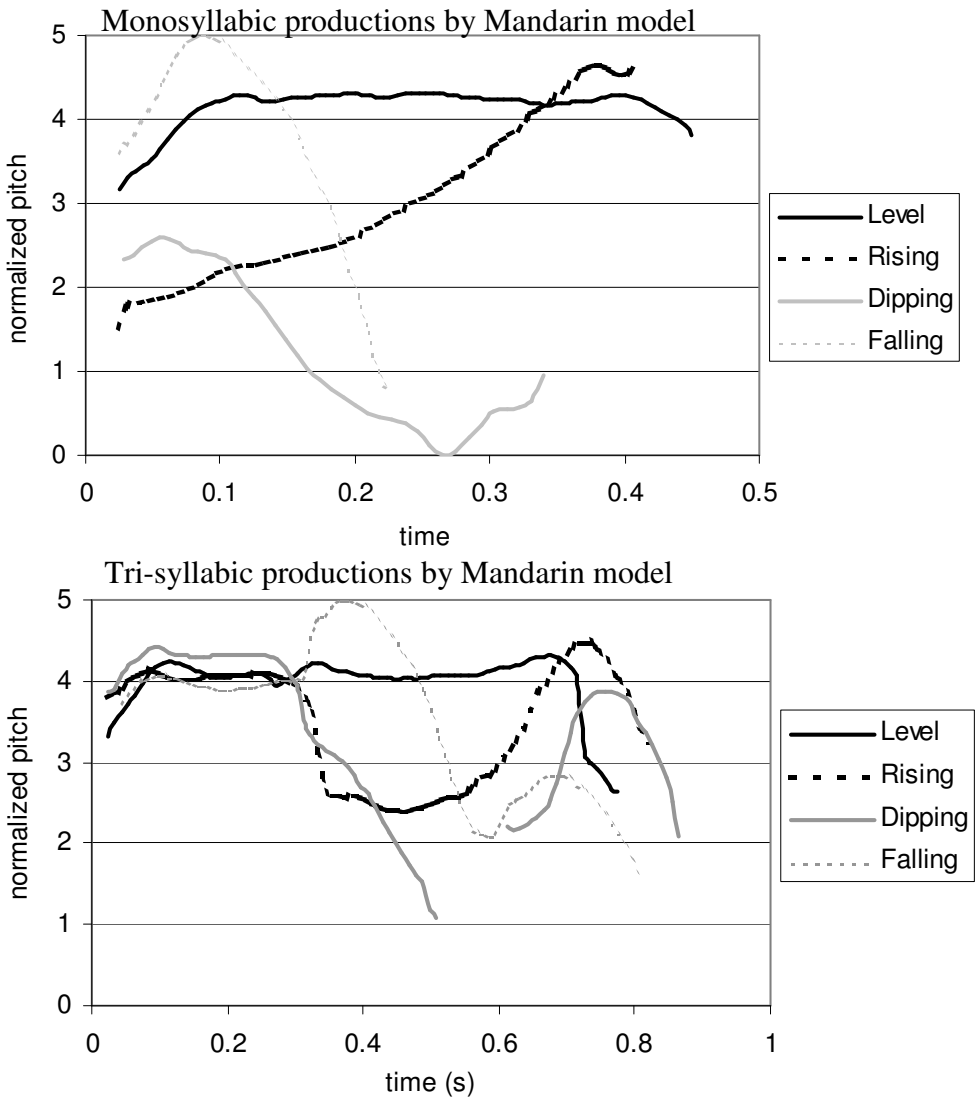


Figure 4.1: The pitch contours for the model Mandarin talker used in the imitation task. The top graph shows the monosyllabic production models and the bottom graph the tri-syllabic production models. The level tone is shown with a solid black line, the rising with a dashed black line, the dipping with a solid gray line, and the falling with a dashed gray line. The break in pitch track for the dipping tone in the tri-syllabic condition indicates that pitch values were unmeasurable due to glottalization.

4.2.3 Task

The participants heard a production model from a loud speaker and were recorded imitating the stimulus. Participants had three seconds to produce their imitation. The production models were blocked by monosyllabic and tri-syllabic production models. The production models were produced in the following order: level, rising, dipping and falling. This series of production models was presented five times. The tri-syllabic production models followed and were presented in the same order. This series of production models was also presented five times. The procedure lasted approximately twenty minutes. The recording procedure and equipment were the same as described for Experiment 1. The instructions given to the participants are shown in Appendix I.

4.2.4 Analysis

Productions were analyzed using native Mandarin listener judgments and acoustic analysis following Wang et al. (2003).

4.2.4.1 Native Mandarin listener judgments

The native English listeners' imitations were submitted to perceptual evaluation by six native Mandarin judges (2 male and 4 female; average age = 27.6, range 25-30). The Mandarin judges listened to all productions by the 20 native English talkers hearing each stimulus just once. The time to complete these judgments (20 talkers x 5 repetitions of each stimulus x 4 monosyllabic stimuli and 4 tri-syllabic stimuli = 800 judgments in all) was approximately 20 minutes. The stimuli were blocked by the number of syllables

(i.e. monosyllables or tri-syllables) but were otherwise randomized. The judge's task was to listen to each production and indicate which tone she or he heard by pressing one of five buttons corresponding to the four Mandarin tones or "none" indicating that the production did not correspond to any of the four Mandarin tone categories. For the three syllable productions, judges were instructed to determine the tonal category for the second syllable in the utterance. They were informed of the intended tones for the first and third syllables in these tri-syllabic utterances. The judges were encouraged to categorize the productions into one of the four tonal categories but the availability of the "none" response allowed for tones which could not be put into any of the four categories (e.g. a level tone in the middle of the pitch range or a rising-falling tone). Before judging the native English talkers' imitations, the Mandarin judges listened to the native Mandarin model's productions hearing each one once. The judges were not required to make any responses during this portion of the experiment. The instructions for the native Mandarin judges are shown in Appendix J.

4.2.4.2 Acoustic Analysis

The productions were submitted to acoustic analysis. The acoustic analysis served to determine the cause(s) of the Mandarin judges' errors and to determine the specific ways in which the native English talkers deviated from the Mandarin model. In order to facilitate comparisons both across the English participants and between the English participants and the Mandarin talker model, the pitch and duration of the tones were

normalized. Duration was normalized by stretching or shrinking the English talkers' productions to match the Mandarin model's durations for each tone. As with the analysis of the stimuli in the perception experiments, all F0 values were normalized per speaker using the following equation:

$$T=[(lgX-lgL)/lgH-lgL]x5$$

where X is the pitch value in hertz, L is the lowest pitch measurement for that speaker, H is the highest measurement for the speaker. The resulting values range from 0 to 5 and correspond to the pitch values in proposed by Chao (1948) to account for the differences across the four Mandarin tones.

For the monosyllabic stimuli, averaged curves were calculated for all 20 talkers based on their five productions of each tone. Additionally, an average curve for each tone for all the talkers was calculated in order to compare the group results to the Mandarin model. For the tri-syllabic stimuli, a subset of the participants were selected for analysis based on their overall d' scores for tri-syllabic stimuli in Experiment 1 and on the accuracy with which their productions were identified by the Mandarin judges. These talkers were selected as a representative sub-set of various types of perception/production relationships including high perception scores and high production scores, high perception scores but low production scores, and low perception scores but high production scores. There were no talkers with both low perception and production scores. These choices were made in order to keep the acoustic analysis manageable while still analyzing a range of participants with different perception and production abilities and varying perception/production relationships.

Pitch measurements were taken for the monosyllabic stimuli based on the average curves for each talker. The pitch values at the onset (0%), 25%, 50%, 75% and offset (100%) points into the syllable and the peak (highest pitch value) and valley (lowest pitch value) were measured. Additionally, overall pitch range and falling and rising range, as appropriate, were calculated. The participants' productions were compared to the Mandarin model, which served as the edge of the expected production space.

4.3. Results

4.3.1 Identification judgments

Overall, the native Mandarin listeners accurately identified the native English participants' productions. On average, the monosyllabic stimuli were identified with 72 percent accuracy and the tri-syllabic stimuli with 73 percent accuracy. Each native English talker's production identification score is shown in Figure 4.2. The correlation between the accuracy with which each talker could imitate the monosyllabic stimuli and the tri-syllabic stimuli based on the identification judgment scores was not significant (Rho = -0.42, p = 0.853).

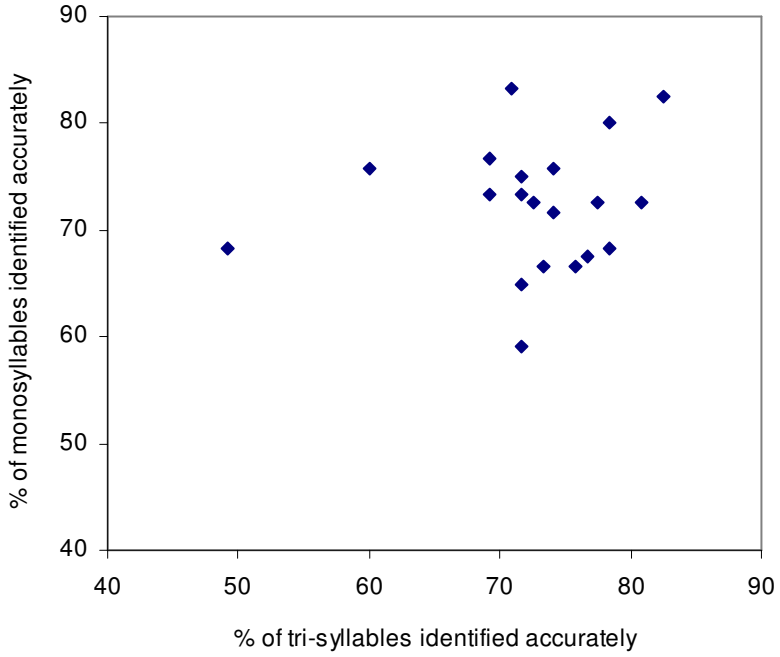


Figure 4.2: The percentage of productions correctly identified by the Mandarin judges for the monosyllabic stimuli (y-axis) and the tri-syllabic stimuli (x-axis) for each native English talker.

While the overall percent correct identifications give a general indication about the accuracy with which the native English talkers were able to imitate the model Mandarin talker, the Mandarin judges' response patterns for each tone provide more information about which particular tones were problematic and the patterns of confusion among the tones. Figure 4.3 and Table 4.1 show the averaged categorization responses for each tone in the monosyllabic and tri-syllabic conditions for the Mandarin judges.

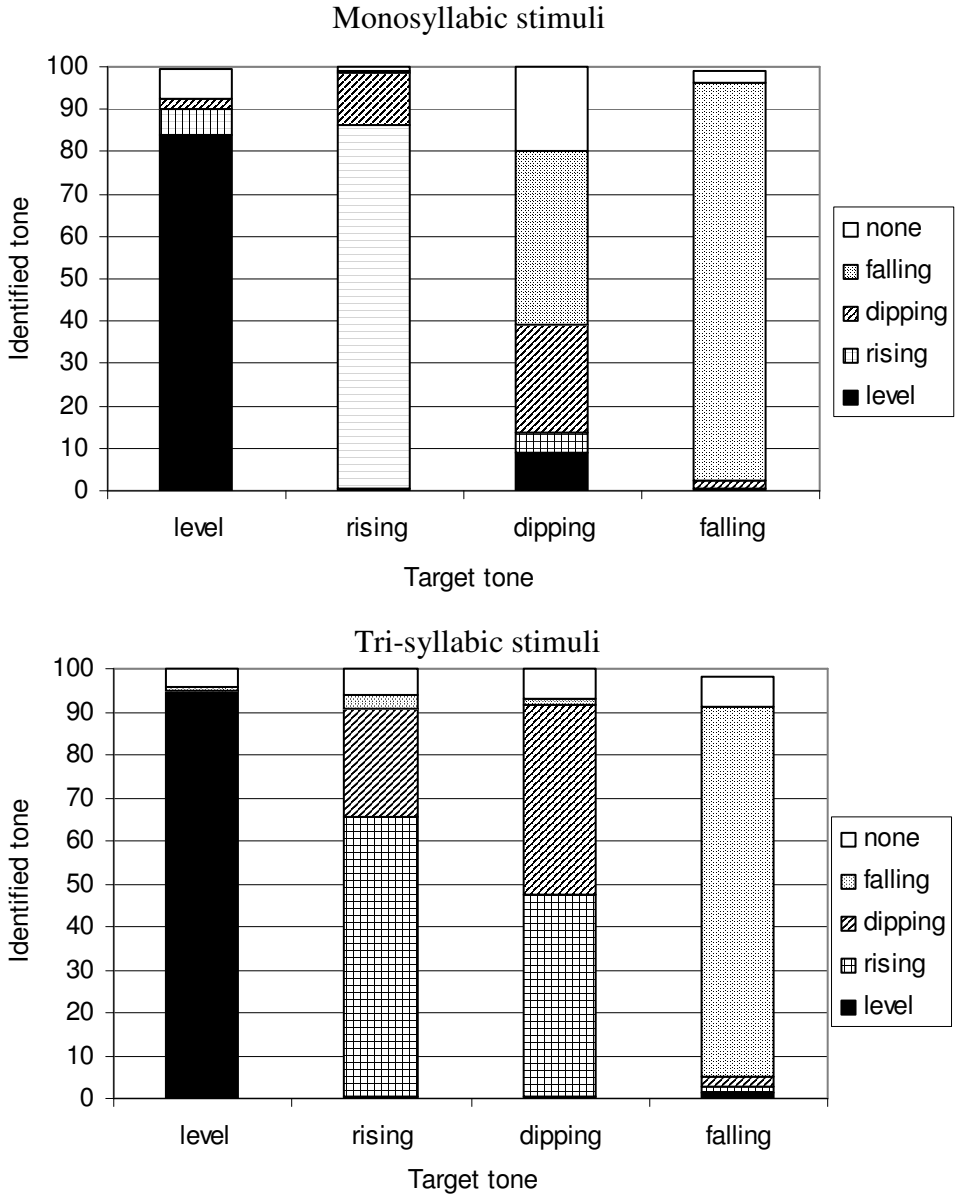


Figure 4.3: The averaged identification judgments by the native Mandarin judges for the native English speakers' productions for the monosyllabic stimuli (top graph) and the tri-syllabic stimuli (bottom graph). The Mandarin judges' identification rates for the level (solid black), rising (cross-hatched), dipping (diagonal lines), falling (dots), and none (white) response categories are shown for each of the four intended tonal targets. The x-axis displays the intended tonal target and the y-axis shows the percentage of

identification judgments for each category (i.e. the four lexical tones and the none category).

Table 4.1: Confusion matrices for the averaged identification judgments by the native Mandarin judges for the native English speakers' productions for the monosyllabic stimuli (top table) and the tri-syllabic stimuli (bottom table). The Mandarin judges' identification rates for the level, rising, dipping, falling, and none response categories are shown for each of the four intended tonal targets.

		Monosyllabic stimuli			
		Target tone			
		Level	Rising	Dipping	Falling
Identified tone	Level	84	0	9	0
	Rising	6	86	5	0
	Dipping	2	13	25	2
	Falling	0	0	41	94
	None	7	1	20	3

		Tri-syllabic stimuli			
		Target tone			
		Level	Rising	Dipping	Falling
Identified tone	Level	95	0	1	1
	Rising	0	65	47	2
	Dipping	0	25	44	2
	Falling	1	3	2	86
	None	4	6	7	7

The identification accuracy scores were analyzed with a repeated measures ANOVA with number of syllables (one versus three) and tone (level, rising, dipping, and falling) as the two within-group factors. The main effect of syllable number was not

significant ($F(1,114) = 0.008$, $p=0.930$), but there was a significant main effect of tone ($F(3, 114)=135.1$, $p<0.0001$) and an interaction of number of syllables and tone ($F(3, 114)=16.5$, $p<0.0001$). The main effect of tone resulted from the overall lower identification accuracy scores for the rising and dipping tones compared to the level and falling tones. For both the monosyllabic and tri-syllabic stimuli, the level and falling tones were identified very accurately. However, the interaction between syllable number and tone is due to the fact that the rising tone was identified much more accurately in the monosyllabic condition than in the tri-syllabic condition while the dipping tone was not identified very accurately in either the monosyllabic or the tri-syllabic condition (although the productions were identified somewhat more accurately in the tri-syllabic condition).

In addition to the overall percent correct identification rates, the examination of the Mandarin judges' patterns of (mis)identifications reveals additional information about native English participants' accuracy and the types of deviations from the Mandarin model. First, for the vast majority of the productions, the Mandarin judges were able to classify the productions into one of the four Mandarin tone categories and used the "none" category sparingly. Second, when the rising tone was misidentified it was most commonly identified as the dipping tone in both the monosyllabic and tri-syllabic conditions. However, the dipping tone was most often misidentified as the falling tone in the monosyllabic condition but as the rising tone in the tri-syllabic condition.

In sum, the Mandarin judges were able to correctly identify the vast majority of the English talkers' imitations in both the monosyllabic and tri-syllabic conditions. The

level and falling tones were identified with high degrees of accuracy in both syllable conditions. There were some differences across the syllable conditions for the rising and dipping tone in terms of accuracy and patterns of confusion. These results suggest that the English speaking participants were able to imitate the Mandarin model with a high degree of accuracy. The productions were next submitted to acoustic analysis in order to identify the acoustic sources of the Mandarin judges' misidentifications and to determine the ways in which the participants' productions matched or deviated from the Mandarin model.

4.3.2 Acoustic analysis

4.3.2.1 Acoustic analysis of monosyllabic productions

Figure 4.4 shows the pitch contours for the four Mandarin tones averaged across the 20 native English participants. As demonstrated by the figure, the participants were able to distinguish the tones quite well and none of the four categories were collapsed. An examination of this graph and the averaged T values for each tone (shown in Tables 4.1 – 4.4) helped to identify the ways in which the native English talkers' productions deviated from the Mandarin model. For most of the tones, the average reflects the performance of most talkers although there was some individual variation observed (see below for discussion of some individual talkers). The only tone in which the average does not accurately reflect the individual performance is the dipping tone. While the

average closely approximates the Mandarin model, individual English participants produced tones that tended to deviate much more from the model.

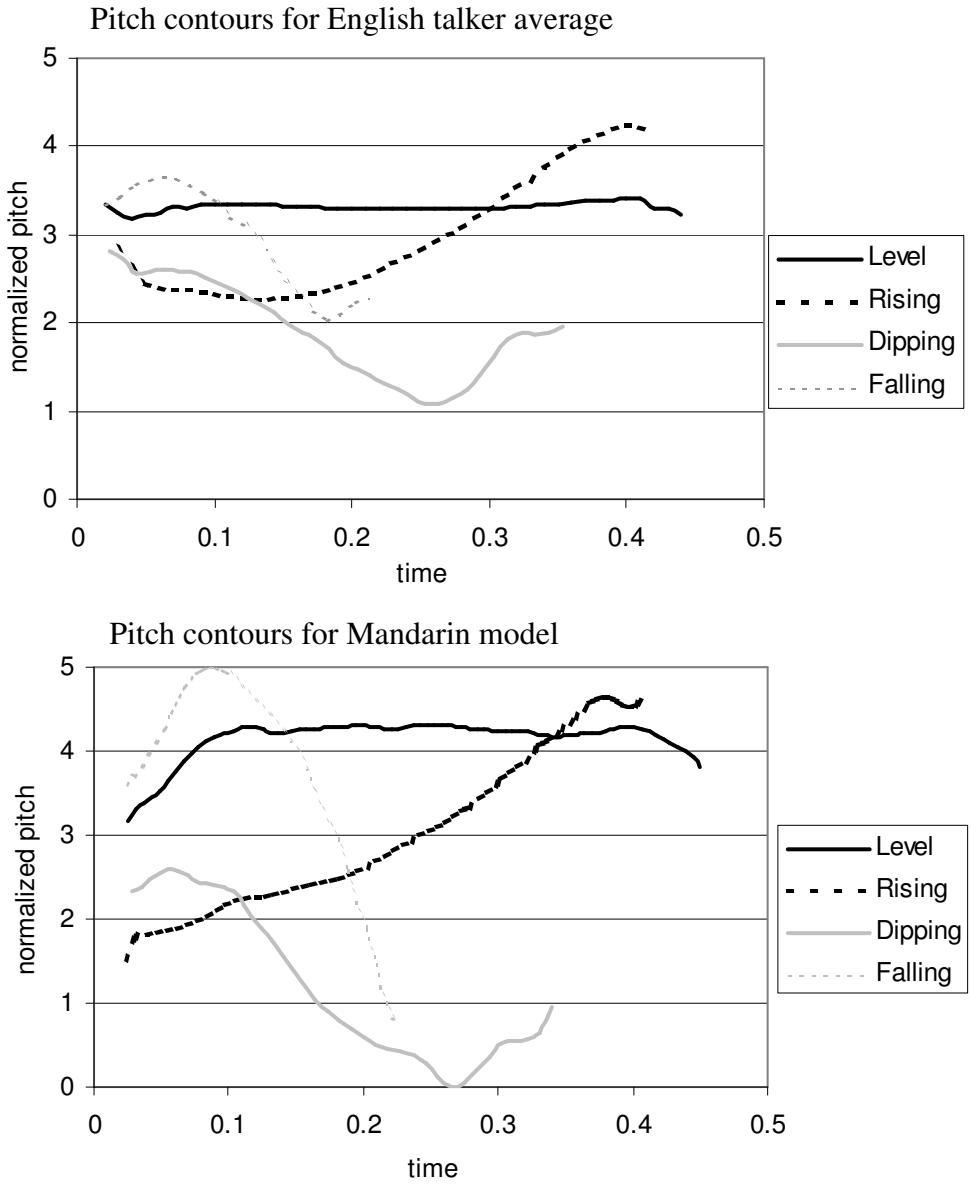


Figure 4.4: The pitch contours for the averaged native English talkers' productions of the four Mandarin tones (top graph) and the pitch contours for the Mandarin model (bottom graph). The level tone is shown in solid black, the rising in dashed black, the

dipping in solid gray and the falling in dashed gray. The x-axis displays the time in seconds and the y-axis shows the normalized pitch values in T-values.

The Mandarin judges very accurately identified the level tones (84% correct). This high degree of accuracy reflects the fact that the native English productions closely matched the Mandarin model. Table 4.2 displays the averaged T values at onset, 25%, 50%, 75% and offset points into the syllable and the pitch range of the pitch contour for the native English talkers compared to the Mandarin model.

Table 4.2: Level tone pitch measurement in T values for the English listeners with standard deviations shown in parentheses and for the Mandarin model.

	0%	25%	50%	75%	100%	Peak	Valley	Range
native English	3.34 (0.84)	3.33 (0.62)	3.29 (0.61)	3.32 (0.60)	3.25 (0.65)	3.69 (0.48)	2.94 (0.72)	0.75 (0.33)
Mandarin model	3.29	4.28	4.26	4.22	4.04	4.32	3.27	1.03

For the level tone, the English listeners deviated from the Mandarin model slightly in terms of pitch height as their productions displayed lower T values throughout the syllable except for the onset. The Mandarin model actually had a larger pitch range due mostly to the initial rising portion of the tone. The difference between the English talkers' productions and the Mandarin model did not substantially hinder the Mandarin judges' abilities to identify these tones particularly since the deviant level tone did not resemble any of the other three tone categories.

The Mandarin judges were also very accurate in their identifications of the rising tone (86% correct) suggesting that the English listeners accurately captured the essential elements of this tone. When the Mandarin judges misidentified the rising tone, they almost always identified it as dipping (13%) with only one percent of responses in the "none" category and no responses for the level or falling categories. Table 4.3 displays the averaged T values at the onset, 25%, 50%, 75% and offset points into the syllable as well as the peak, valley, and rising range (calculated from valley to offset).

Table 4.3: Rising tone pitch measurements (expressed in T values) for the English listeners with standard deviations shown in parentheses and the Mandarin model.

	0%	25%	50%	75%	100%	Peak	Valley	Range	Rising range
native English	2.82 (.93)	2.27 (0.75)	2.57 (0.64)	3.51 (0.57)	4.14 (0.54)	4.32 (0.42)	2.07 (0.79)	2.25 (0.54)	2.07 (0.42)
Mandarin model	1.82	2.23	2.66	3.46	4.60	4.62	1.82	2.79	2.78

For the rising tone, the English listeners' productions displayed a more compressed pitch range than the Mandarin model resulting from onset and offset pitch values that were higher and lower, respectively, compared to the Mandarin model. However, the overall shape of the pitch contour displayed a rising pattern and was approximately in the correct region of the pitch range.

Figure 4.5 displays the pitch contours for the Mandarin model, the average of the English talkers and for the native English talker whose rising tones were identified least accurately. For the talker whose imitations were identified poorly, the rising tones were

identified as rising 67% and identified as dipping the other 33%. From an inspection of the talker's pitch contours, it is clear that the overall shape of the contour most likely lead to the misidentifications as the talker produced a significant falling portion before the rising portion at the end of the syllable. However, this talker did produce the contour in the appropriate section of the pitch range.

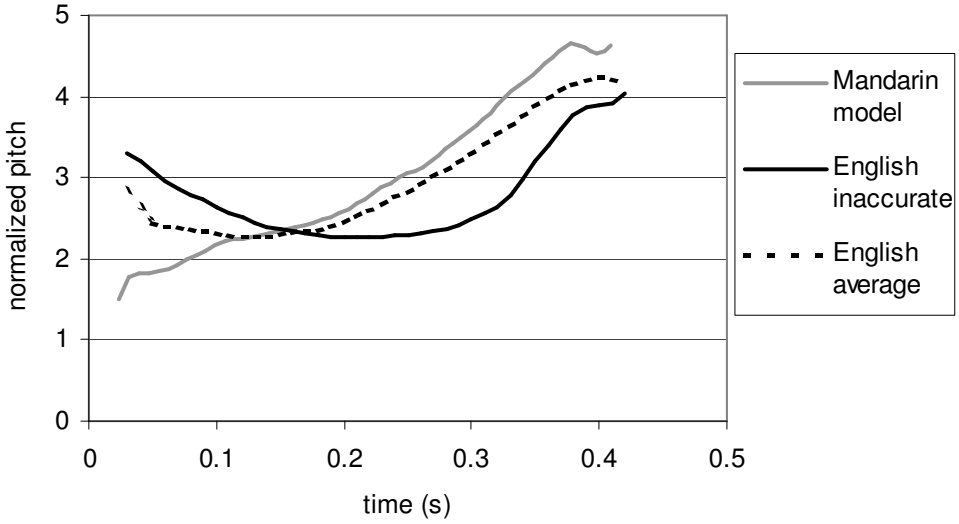


Figure 4.5: Pitch contours for the Mandarin model (solid gray line), the average of the English talkers (dashed black line), and one individual English talker (solid black – English inaccurate) whose productions of the rising tone the Mandarin judges often identified as dipping.

For the dipping tone, the Mandarin judges were unable to accurately identify the majority of the native English imitations. Only 25% of the production were labeled correctly while 41% were labeled as falling, 20% as "none", 9% as level, and 5% as

rising. Table 4.4 displays the average T values for the native English talkers and for the Mandarin model. The values at the onset, 25%, 50%, 75% and offset points into the syllable are included as well as the peak, valley, pitch range, the falling range (calculated from onset to valley) and the rising range (calculated from valley to offset).

Table 4.4: Dipping tone pitch measurement (expressed in T values) for the English listeners with standard deviations shown in parentheses and the Mandarin model.

	0%	25%	50%	75%	100%	Peak	Valley	Range
native English	2.80 (0.64)	2.45 (0.70)	1.70 (0.83)	1.17 (0.58)	1.81 (1.09)	2.96 (0.66)	0.91 (0.63)	2.05 (0.77)
Mandarin model	2.43	2.40	0.92	0.21	0.55	2.58	0.01	2.57
	Rising range	Falling range						
native English	0.90 (0.98)	2.07 (0.67)						
Mandarin model	0.55	2.41						

For the dipping tone, on average the English talkers did not produce falls as extensive as the Mandarin model and displayed pitch values that were not as low as the Mandarin model during the middle and end of the syllable. Because of the widely variable responses of the Mandarin judges, it is informative to look more closely at particular talkers who displayed the various patterns of misidentification. The three talkers shown below in Figure 4.6, contrasted with the Mandarin model, had dipping tones that were frequently identified as falling or “none”. Two of the talkers’ (talkers “English falling 1” and “English falling 2”) dipping imitations were identified as falling

73% of the time and 80% of the time, respectively. The other talker shown in Figure 4.6 ("English none") produced dipping tone imitations that were frequently labeled as "none" (37%). The remaining portion this talker's dipping tones were identified correctly (40%), as the falling tone (17%), or as the rising tone (7%).

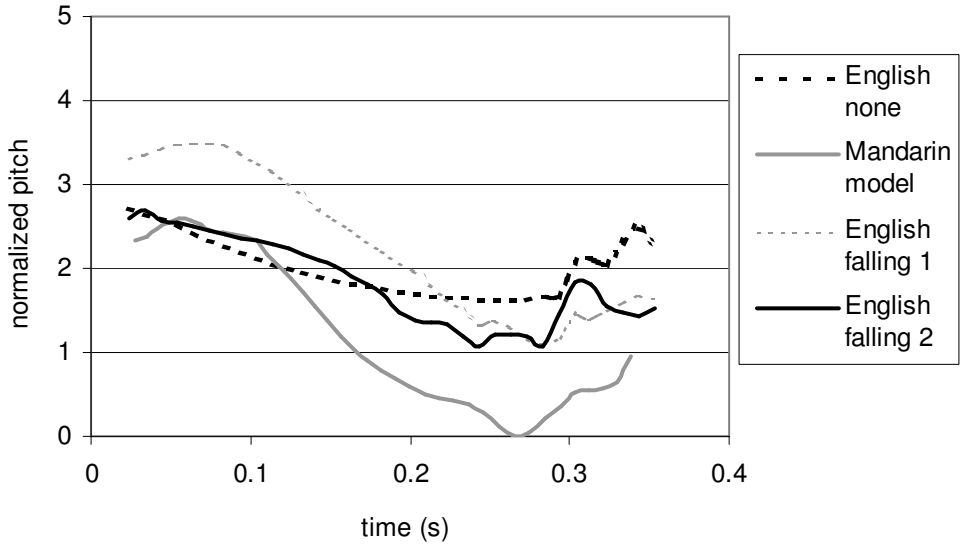


Figure 4.6: Dipping pitch contours for three native English talkers and the Mandarin model (solid gray line). English falling 1 (dashed gray line) and English falling 2's (solid black line) dipping tone imitations were predominantly identified as falling. English none's (dashed black line) dipping tone imitations were frequently identified as not fitting into any of the four Mandarin tone categories.

Talker English falling 1 whose imitations were frequently labeled as falling started her productions higher than the Mandarin model's production. Her starting value for the dipping tone is closer to the Mandarin model's onset pitch value for the falling tone. Both English falling 1 and 2's lowest pitch values were not as low as the Mandarin

model. Furthermore, while not represented in the pitch contour graphs, both of these talkers had a glottalized voice quality at the end of their syllables for three out of the five repetitions. When the dipping tone is glottalized, it tends to be glottalized in the middle of the syllable whereas the falling tone displays glottalization at the end of the syllable. This voice quality factor, in addition to the pitch height differences compared to the Mandarin model, may have lead the Mandarin judges to perceive the dipping tones as falling. The cause of the frequent "none" identification for Talker English none may have been the compressed pitch range (1.09) compared to the Mandarin model (2.79) and the fact that lowest pitch value for this talker was only 1.61 compared to the much lower value of the Mandarin model (0.01). All of the above deviations can be related to pitch height and range suggesting that even for this difficult to imitate tone, the participants were accurately able to imitate the contour shape but the height and range deviations lead to the misidentifications by the Mandarin judges.

For the falling tone, the Mandarin judges were able to very accurately identify the English talker's productions with 94% correct and only a few (3%) placed in the "none" category and a few in the dipping category (2%). Table 4.5 displays the average T values for the native English talkers for the onset, 25%, 50%, 75% and offset points into the syllable as well as the peak, valley, overall range, and falling range (calculated from peak to offset).

Table 4.5: Falling tone pitch measurements (expressed in T values) for the English listeners with standard deviations shown in parentheses and the Mandarin model.

	0%	25%	50%	75%	100%	Peak	Valley	Range	Falling range
native English	3.34 (0.74)	3.64 (0.59)	3.25 (0.66)	2.40 (0.93)	2.18 (1.03)	3.73 (0.55)	1.77 (0.75)	1.96 (0.59)	1.57 (0.76)
Mandarin model	3.63	4.88	4.64	3.24	0.98	4.99	0.98	4.01	2.65

For the falling tone, the English talkers displayed a more compressed pitch range than the Mandarin model caused by a lower peak and a higher ending pitch point compared to the Mandarin model. However, the native English productions generally followed a falling pattern and therefore did not resemble any of the other four Mandarin tones. These deviations from the Mandarin model did not seem to substantially hinder the Mandarin judges abilities to correctly identify the syllable as having the falling tone.

4.3.2.2 Acoustic Analysis of Tri-Syllabic Productions

Figure 4.7 displays the pitch contours for the four Mandarin tones in the tri-syllabic context averaged across 7 native English participants and the Mandarin model. These native English talkers were chosen to represent a range in tone production performance (as determined by the Mandarin judge's scores) as well as to represent different relationships between perception and production. As demonstrated by the figure, the participants were able to produce the level and falling tones in the tri-syllabic utterances quite accurately, but the rising and dipping tones showed very similar pitch

contours. Examining specific talkers' productions of the four tri-syllabic tones will demonstrate the ways in which individuals adhered or deviated from the Mandarin model.

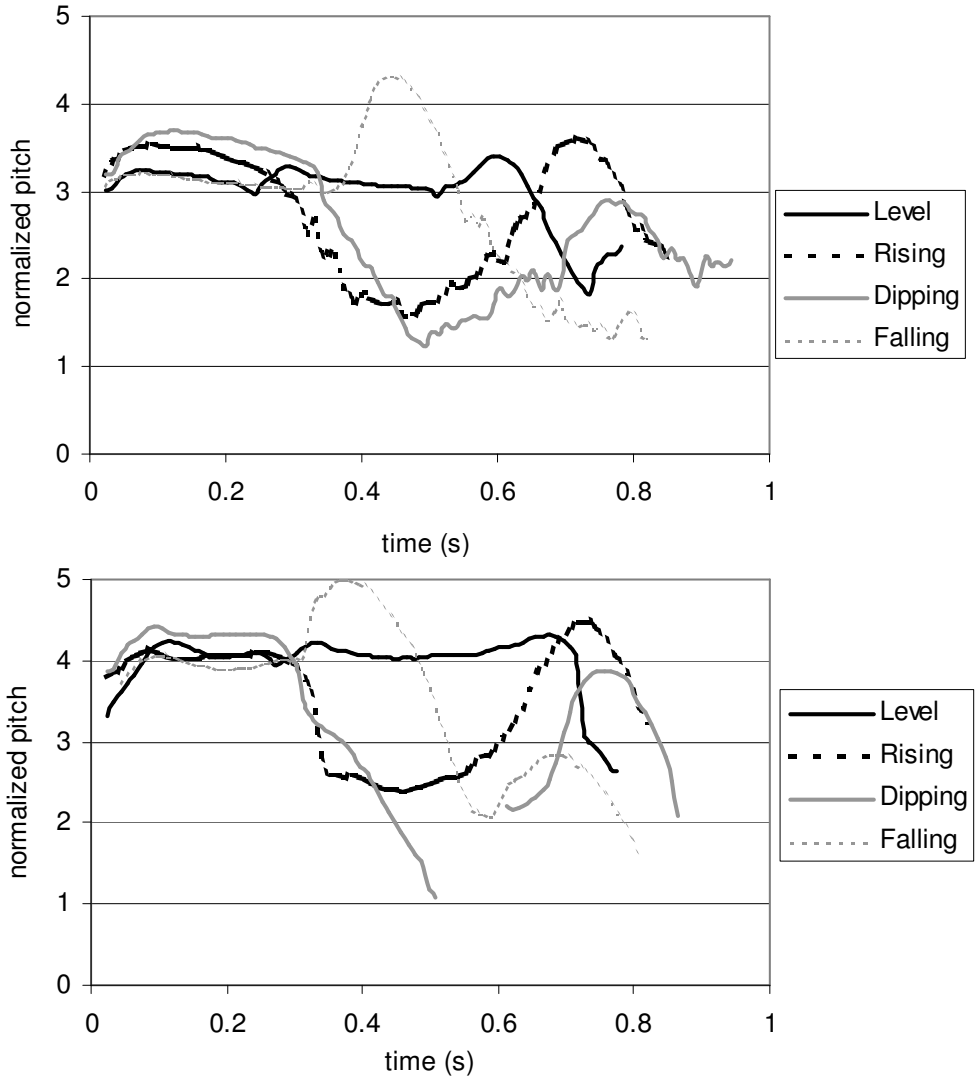


Figure 4.7: The averaged pitch contours of the four Mandarin tones in the tri-syllabic utterances for seven native English talkers (top graph) and the Mandarin model (bottom graph). The level tone is shown in solid black, the rising in dashed black, the dipping in solid gray, and the falling in dashed gray.

The level tone was overall identified very accurately for all 20 native English talkers. Across all participants, the level tone was very accurately identified (95%

correct). For the subset of seven talkers, the level tone was identified with 92% accuracy. Two talkers were selected to represent the edges of this accuracy distribution for comparison with the Mandarin model. Shown below in Figure 4.8 are the level tone pitch contours in the tri-syllabic utterances for the Mandarin model, the average of the seven native English talkers, and two individual native English talkers. One native English talker's productions (talker s50) were identified with 100% accuracy while the other talker's productions (talker s60) were identified correctly 77% of the time with the remaining productions identified as rising (10%) or as none (13%). Based on visual comparison of the pitch contours, for the average English talkers' pitch contour and the contours for the two individual talkers, the general shape of the contours are very similar to the Mandarin model. However, the native English average pitch contour is lower in the pitch range than the Mandarin model. The talker whose productions were identified with perfect accuracy had a pitch contour that is much closer to the Mandarin model in pitch range than the talker whose productions were identified less accurately.

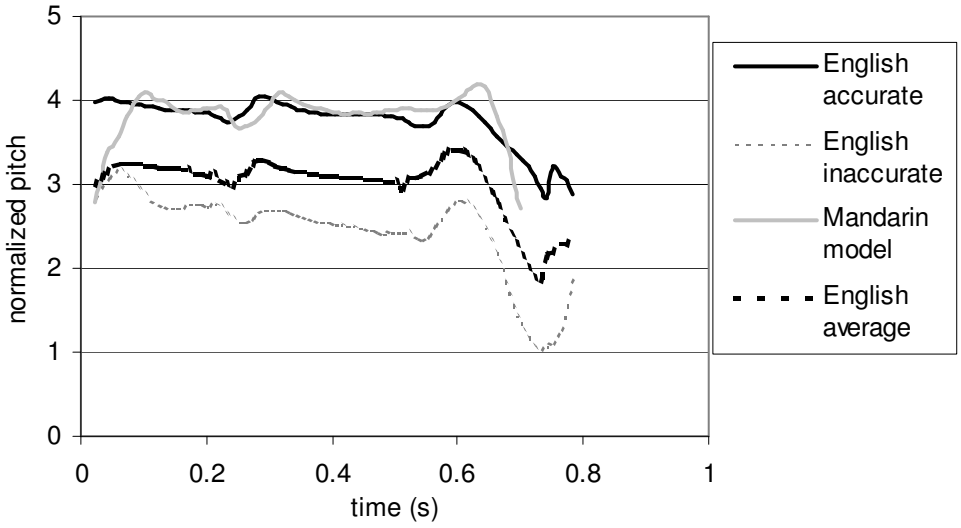


Figure 4.8: Tri-syllabic level pitch contours for the seven English talkers analyzed (black dashed line), the Mandarin model (gray solid line) and two specific talkers who represented extremes of identification accuracy scores by the Mandarin judges. Talker English accurate's productions (black solid) were identified perfectly while talker English inaccurate productions (dashed gray line) were identified less accurately.

While the level tone was identified very accurately, the rising and dipping tones were identified accurately much less frequently and were often confused with one another. This pattern is not surprising considering that the Mandarin listeners appear to be less sensitive to this contrast in this particular tri-syllabic context even with native Mandarin talkers (see findings from Experiments 1 and 2). However, the Mandarin listeners were able to discriminate the productions from the Mandarin model with 97% accuracy in Experiment 1. Furthermore, when the Mandarin judges, who identified the native English productions, were asked to identify the productions by the Mandarin

model talker, they were able to identify the productions of the rising and dipping tones with 100% accuracy. Therefore, if the English talkers accurately imitated the Mandarin model, the Mandarin judges should have been able to accurately label the two tones.

The rising tone was identified for all 20 talkers with 65% accuracy and frequently misidentified as the dipping tone (25%). For the subset of the talkers whose productions were acoustically analyzed, the rising tone was identified with 64% accuracy and misidentified as the dipping tone 31% of the time. The dipping tone was identified even less accurately (44% correct for all 20 talkers and 54% for the seven talkers acoustically analyzed) and was frequently identified as the rising tone (47% for all 20 talkers and 37% for the seven talkers acoustically analyzed). In Figure 4.9, the pitch contours for rising and dipping tones for the Mandarin model and the average of the seven English talkers are shown.

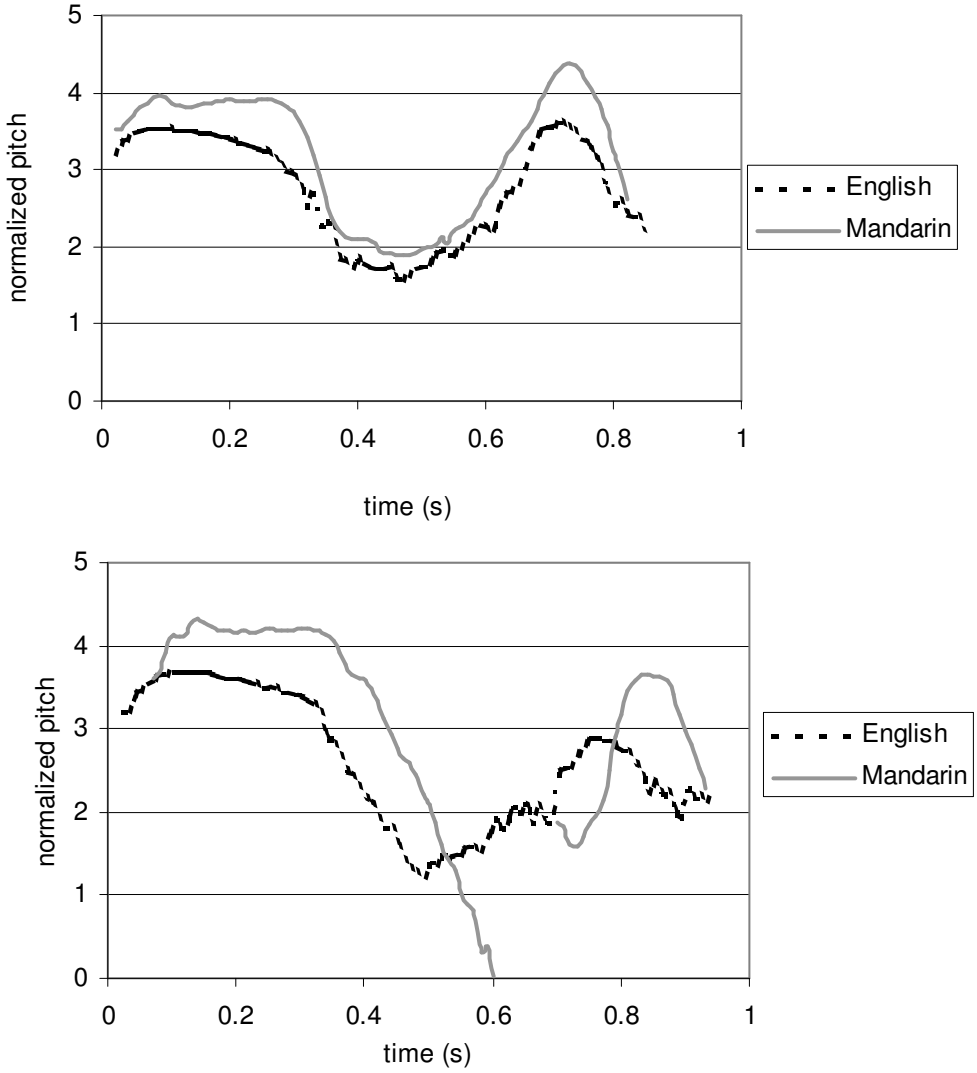


Figure 4.9: Pitch contours for the rising tone (top graph) and the dipping tone (bottom graph) in the tri-syllabic utterances. The average of the seven analyzed English talkers (dashed black line) and the Mandarin model (solid gray line) are shown.

As with the level tone, it appears that the English talkers were able to accurately imitate the general shape of the rising and dipping contours but produced the majority of the two contours in a slightly lower pitch range. For the rising tone, the whole contour

was shifted to a lower pitch range. For the dipping tone, the middle section of the contour was in approximately the same pitch range as the Mandarin model although the valley was not as low but the first and last third of the contour were in a lower region of the pitch range. Additional factors not represented in the pitch contour graphs could have contributed to the misidentification of these two tones including lack of glottalization for the dipping tone or glottalization of the rising tone. These voicing quality differences seem to be one of the key distinguishing features for the Mandarin model. Therefore, inaccurate imitations of the voice quality differences between the rising and dipping tones could have lead to the Mandarin judges' confusions.

For all 20 talkers, the falling tone in the tri-syllabic utterances was identified very accurately (86%) with 7% of the production identified as in the "none" category and only 2% identified as falling, 2% as dipping, and 1% identified as level. For the subset of seven talkers, 78% of the productions were correctly identified with 11% identified in the "none" category and 3% identified as level, 3% as dipping, and 5% as rising. In Figure 4.10, the pitch contours of the average of the seven analyzed English talkers, the Mandarin model, and two individual English talkers who represent the extremes of identification accuracy as indicated by the Mandarin judges are shown. Talker English accurate's production were always identified as falling whereas talker English inaccurate's production were only identified as falling 20% with the other identification judgments distributed among the other categories (20% as level, 23% as rising, 7% as dipping and 30% as none).

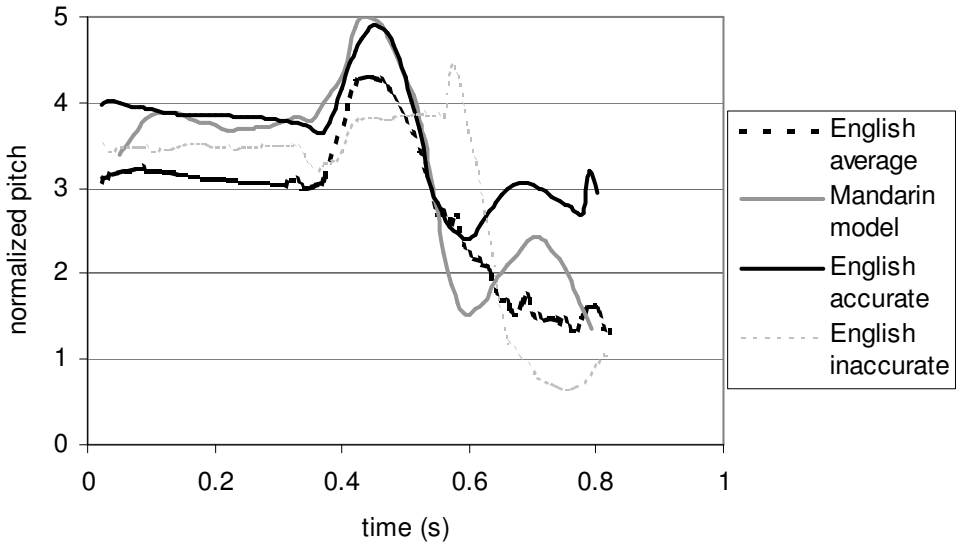


Figure 4.10: Pitch contours for the falling tone in the tri-syllabic condition for the seven English talkers analyzed (black dashed line), the Mandarin model (gray solid line) and two individual talkers who represented extremes of Mandarin judges identification accuracy scores. Talker English accurate's productions (black solid line) were identified perfectly while talker English inaccurate's productions (dashed gray line) were identified very inaccurately.

Overall the shape of the seven English talkers closely resembled the Mandarin model although the initial part of the utterance was produced in a lower pitch range and the peak in the utterance was lower. Furthermore, the average contour did not display a rise before the final falling portion but rather falls continually from the peak in the middle of the utterance. The talker whose productions were identified perfectly very closely resembled the Mandarin model for the first two thirds of the utterance while the final

rise-fall portion of the contour was in a higher region of the pitch range than the Mandarin model. In contrast, the talker whose productions were identified very poorly had a small rise and then a plateau when the Mandarin model was rising to the peak and this talker's fall was delayed without an additional rise before the final fall. These deviations resulted in a contour that did not closely resemble the falling tone but also did not resemble any of the other three Mandarin tones thus leading to the Mandarin judges' inaccurate and highly variable categorization judgments.

4.3.3 Production-perception relationship

Since there was no correlation between the accuracy with which the English participants could imitate the monosyllabic and tri-syllabic Mandarin utterances (Rho = -0.042, p= 0.853), the relationship between production and perception will be investigated separately for the monosyllabic and tri-syllabic stimuli.

4.3.3.1 Production-Perception Relationship Averaged across Tone Pairs

Overall, there was no correlation between the individual participants' ability to accurately produce the monosyllabic stimuli (as measured by the mean accuracy score of the Mandarin judges) and their ability to discriminate the monosyllabic stimuli (as measured by their mean d' score in Experiment 1). In Figure 4.11 the production and perception scores for the 20 participants are shown. The correlation between the measures was not significant (Pearson rank correlation: Rho = 0.142; p = 0.537).

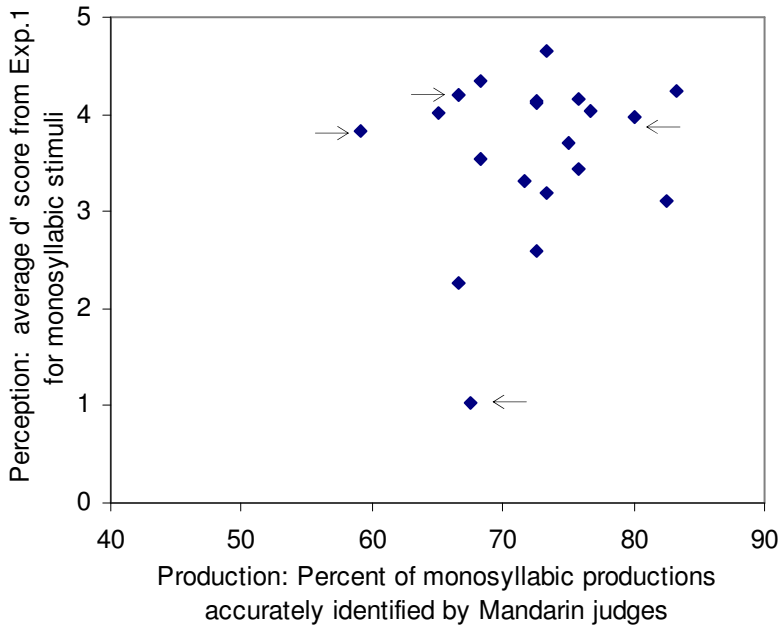


Figure 4.11: The monosyllabic production and perception scores for the 20 English participants. The perception score is the d' score averaged across tone pairs for the monosyllabic stimuli in Experiment 1. The production score is the percent of each participant's monosyllabic productions that the Mandarin judges were able to identify accurately. Arrows indicate the participants who will be discussed in depth below.

The same lack of a relationship between the perception and production scores of individual listeners was found with the tri-syllabic stimuli (Rho = 0.097; p = 0.673). Measurements of production and perception were the same and the individuals' scores are shown below in Figure 4.12.

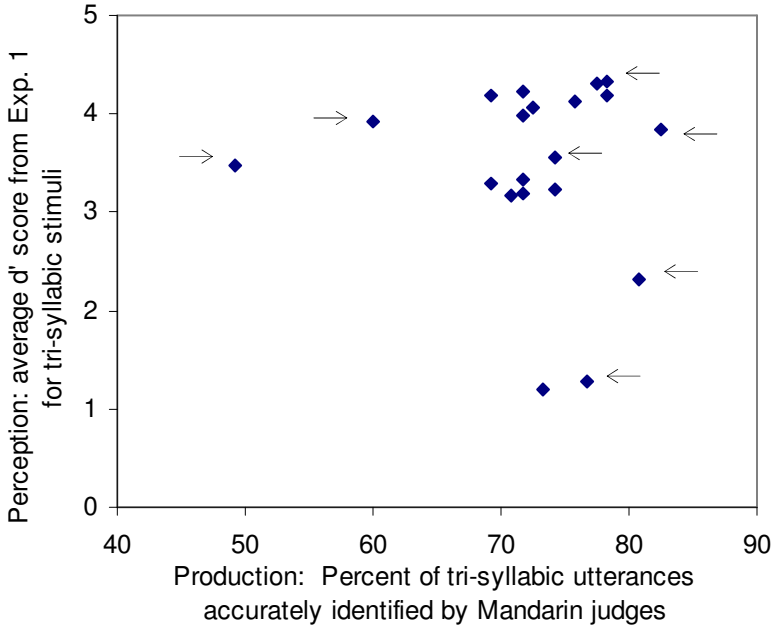


Figure 4.12: The tri-syllabic production and perception scores for the 20 English participants. The perception score is the d' score averaged across tone pairs for the tri-syllabic stimuli in Experiment 1. The production score is the percent of each participant's tri-syllabic productions that the Mandarin judges were able to identify accurately. The arrows indicate which participants productions were acoustically analyzed.

The lack of a correlation between overall perception and production scores for naïve non-native suggest that abilities in perception and production are not related. While there were no participants who were poor on both perception and production, there were participants who were very accurate perceivers while being poor at production,

participants with very accurate production and poor perception, and participants with both high perception and production scores.

4.3.3.2 Relationship between Perception and Production Accuracy for Tone Pairs

There was not a strong relationship between tone pairs that were difficult in perception and those that were difficult in production. The perception and production scores for the tone pairs are shown below in Table 4.6.

Table 4.6: Average percent of tone confusions in production and average d' scores for the discrimination test in Experiment 1. Sensitivity scores are shown for each tone pair and tone confusion scores are calculated based on the percent of times that each target tone within the pair was misidentified as the other tone. For example, the percent of confusions for the level-rising tone pair is an average of the percent of times the Mandarin judges labeled the level tone as rising and the rising tone as level.

	Monosyllabic stimuli		Tri-syllabic stimuli	
	Production	Perception (d')	Production	Perception (d')
level-falling	0.2	3.7	1.1	3.8
rising-falling	0.2	3.5	2.2	3.5
level-rising	3.2	3.2	0.3	3.6
level-dipping	5.6	4.0	0.3	3.8
rising-dipping	8.7	3.6	36.1	2.5
dipping-falling	21.4	3.6	1.9	3.6

Rank-order correlations between the perception and production scores for each tone pair in the monosyllabic and tri-syllabic conditions were calculated. There was no correlation between perception and production for the pairs of monosyllabic stimuli (Rho

= 0.147, p=0.74). The correlation for the tri-syllabic stimuli was not significant but was approaching significance and in the expected direction (Rho = -0.806, p=0.07). A negative correlation would be expected if there was a correlation between perception and production abilities since better production scores are lower (fewer misidentification between the members of the pair) while better perception scores are higher (higher sensitivity to the members of the pair). These non-significant correlations could partially be caused by ceiling effects since the participants performed very accurately in perception except for the rising-dipping pair in the tri-syllabic condition.

For the monosyllabic stimuli, the participants were perceptually very sensitive to all tone pairs. In production, the participants had trouble distinguishing the falling and dipping tones, however. This finding may indicate an instance in which perception leads production. For the tri-syllabic stimuli, the participants were highly perceptually sensitive to all tone pairs except the rising-dipping pair, which received a much lower d' score. The rising-dipping pair is also the pair that the participants had most trouble distinguishing in production. This result indicates a case in which the lack of perceptual acuity leads to an inability to distinguish the pairs in production.

4.3.3.3 Individuals' Perception and Production Accuracy for Tone Pairs

A more thorough investigation of several pairs of talkers who were matched on either their production or perception abilities but displayed very different scores in the other modality is given below. This analysis will serve to determine if, for individual

talkers, the abilities in production and perception are linked. The first two pairs of talkers' perception and production scores for the monosyllabic stimuli are discussed while the second two pairs provide an analysis of the tri-syllabic stimuli.

4.3.3.3.1 Four participants' perception and production accuracy for monosyllabic tone pairs.

The pitch contours for the two talkers with the most widely divergent monosyllabic production abilities while having similar monosyllabic perceptual scores are shown below in Figure 4.13. These two talkers productions were identified with approximately the same moderately accurate scores (66.7% for talker s29 and 67.5% for talker s19) while their perception scores varied widely (d' scores of 4.2 for s29 and 1.0 for s19). These two talkers sensitivity scores and production confusion scores are shown in Figure 4.14.

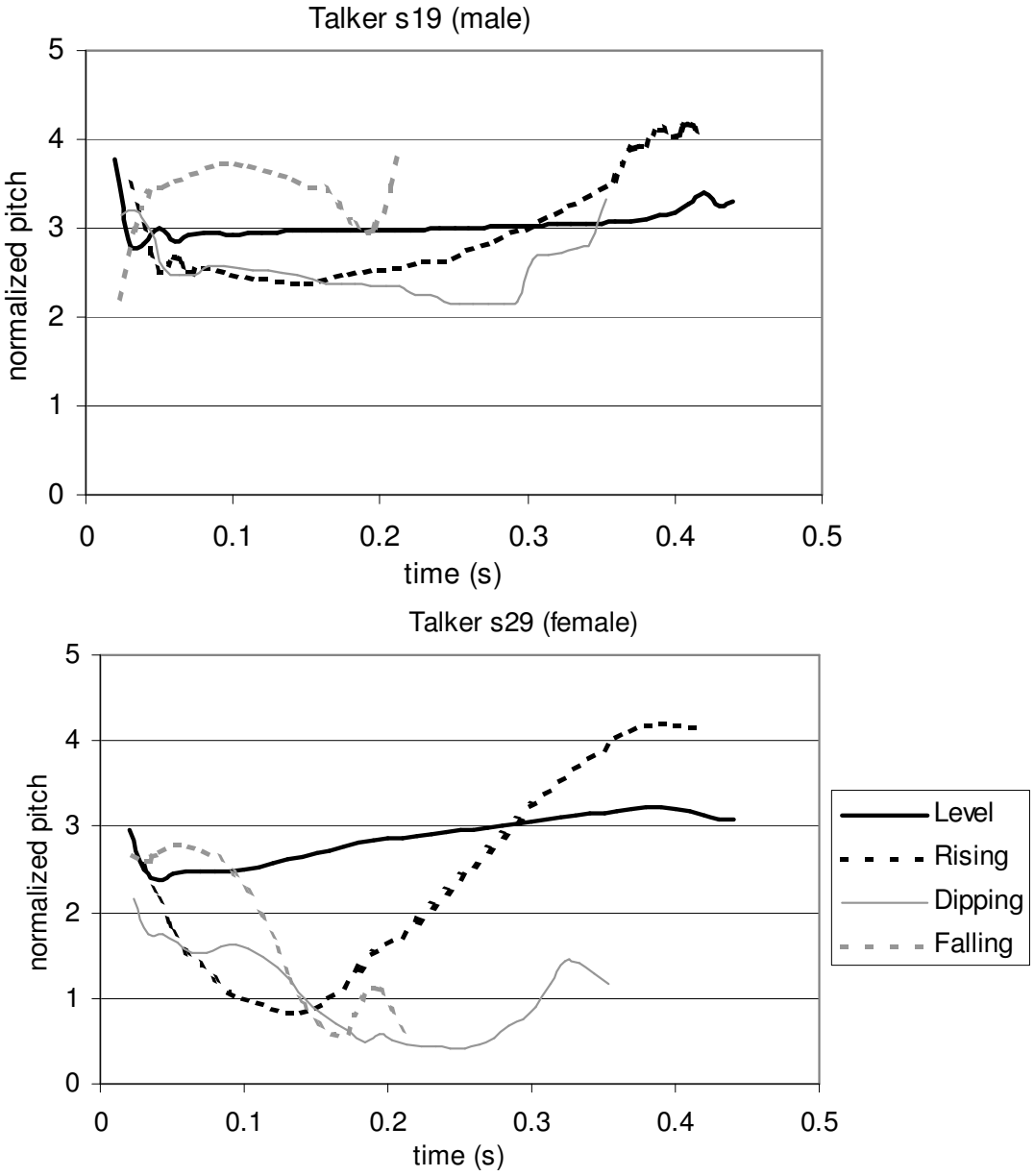


Figure 4.13: Pitch contours for talkers s19 (shown on top) and s29 (shown on bottom) who had very similar monosyllabic production scores but highly divergent monosyllabic sensitivity scores.

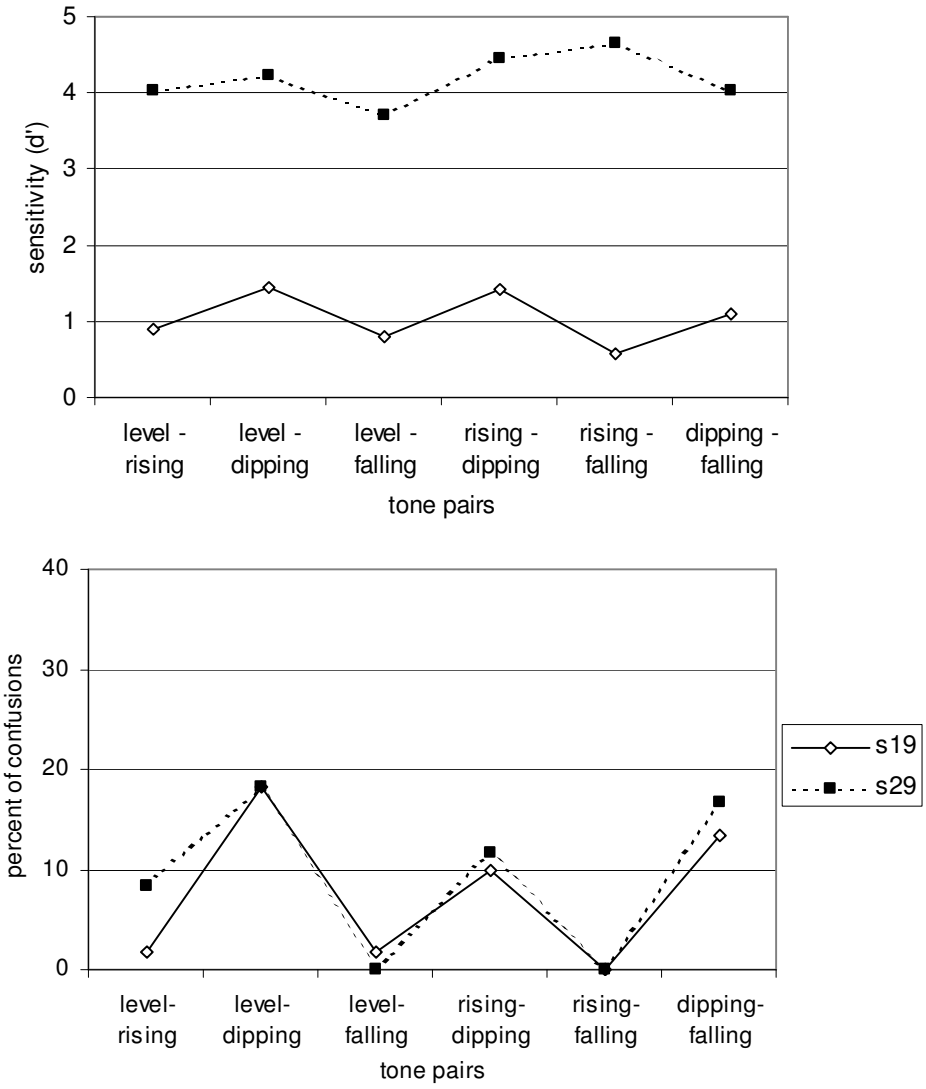


Figure 4.14: Sensitivity scores (shown on top) and production confusion scores (shown on bottom) for two participants, s19 and s29. Sensitivity scores are shown for each monosyllabic tone pair and tone confusion scores also shown for each tone pair are calculated based on the percent of times that each target tone within the pair was misidentified as the other tone. For example, the percent of confusions for the level-rising tone pair is an average of the percent of times the Mandarin judges labeled the level tone as rising and the rising tone as level.

Participant s19, the participant with lower sensitivity, scores displays an inverse relationship between difficulty in perception and production as the tones he could distinguish best in production were those with the lowest sensitivity scores. However, because all his d' scores were very low, it is difficult to make any definitive assessment. This participant demonstrates an instance in which tone pairs which were not well distinguished perceptually were well distinguished in production. Interestingly, although both of these talkers had very similar production scores, talker s19 used a very compressed pitch range compared with talker s29. Therefore, there may be factors other than the Mandarin judges' ability to identify the productions that correlate with perceptual abilities. The ability to consistently use a wide pitch range should be investigated further.

The other two participants whose production and perception scores for the monosyllabic stimuli will be presented in depth show the opposite relationship as the two participants discussed above. These two talkers, s24 and s59, both had very high monosyllabic d' scores (3.8 and 4.0, respectively) but displayed very different monosyllabic production imitation abilities as determined by accurate identification by the Mandarin judges (59.2 and 80.0 correct, respectively). The pitch contours for these two talkers are shown below in Figure 4.15. In Figure 4.16, these two talkers' sensitivity scores and production confusion scores are shown.

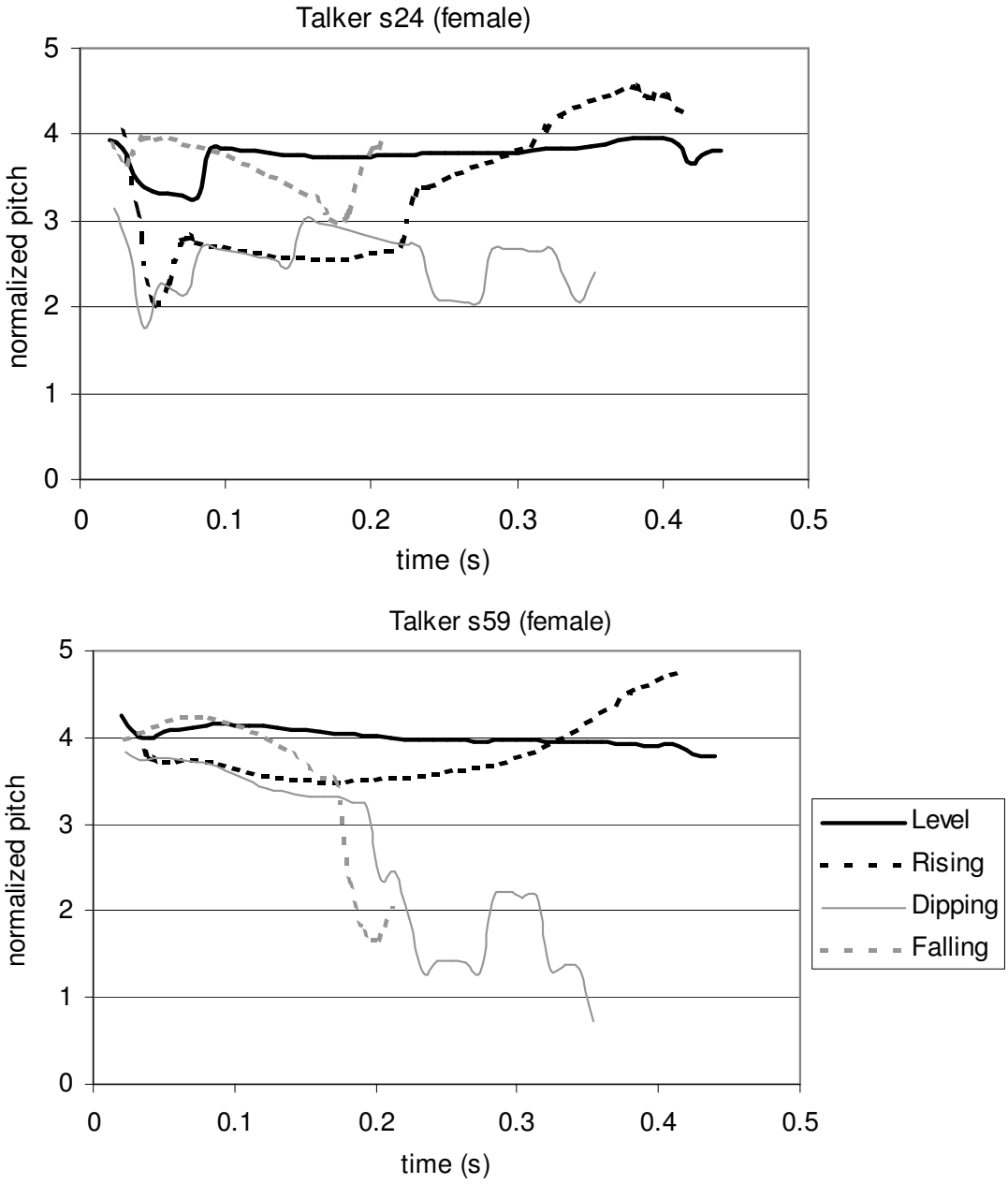


Figure 4.15: Pitch contours for talkers s24 (shown on top) and s59 (shown on bottom) who had very similar monosyllabic sensitivity scores but highly divergent monosyllabic production scores.

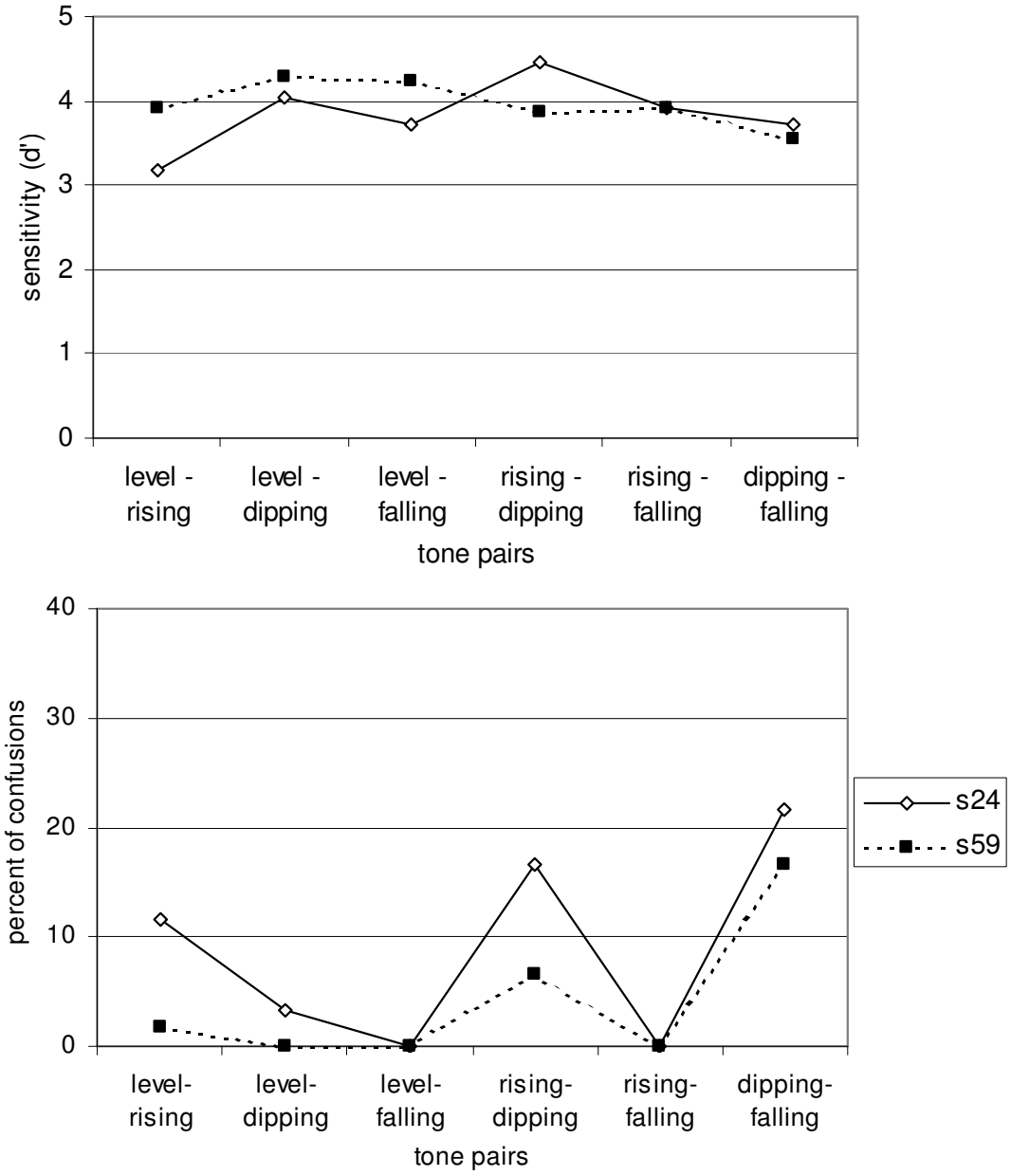


Figure 4.16: Sensitivity scores (shown on top) and production confusion scores (shown on bottom) for two participants, s24 and s59. Sensitivity scores are shown for each monosyllabic tone pair and tone confusion scores are calculated based on the percent of times that each target tone within the pair was misidentified as the other tone.

This pair of talkers demonstrates an instance in which accurate perception does not always lead to accuracy in production. While both of these participants perceptually distinguished all tone pairs very well, one was substantially less accurate in the production task. Both participants had difficulty with the rising-dipping and dipping-falling contrasts in production. In perception, these two tone pairs did not receive substantially lower d' scores. In fact, talker s24's second worst production score (for the rising-dipping pair) corresponded to her highest sensitivity score. Talker s24 had difficulty with the level-rising contrast in production while talker s59 did not. This discrepancy between the two talkers for this pair is mirrored in the perception scores since talker s24 had a lower d' score on the level-rising tone pair than talker s59 did.

From the investigation of these four participants who covered the full range of perception and production abilities and relationships between the two, there is little evidence for a perception-production link for monosyllabic non-native prosodic categories. The perceptual sensitivity scores on particular tone pairs did not seem to predict production accuracy or vice versa.

4.3.3.3.2 Four participants' perception and production accuracy for tri-syllabic tone pairs

The seven talkers who were selected for acoustic analysis on the tri-syllabic stimuli were chosen to represent different types of perception/production relationships. Four of these participants will be discussed in depth below. Two talkers who performed

very accurately on the production task but very differently on the perception task and two talkers who both performed very accurately on the perception task but performed very differently on the production task will be discussed.

The first pair of participants, s19 and s34, to be discussed for the tri-syllabic stimuli both had high production scores (78.3 and 76.7, respectively) but widely divergent perception scores (d' scores of 4.3 and 1.3, respectively). These two talkers' pitch contours for the tri-syllabic stimuli are shown below in Figure 4.17. In Figure 4.18, these two talkers' sensitivity scores and production confusion scores are shown.

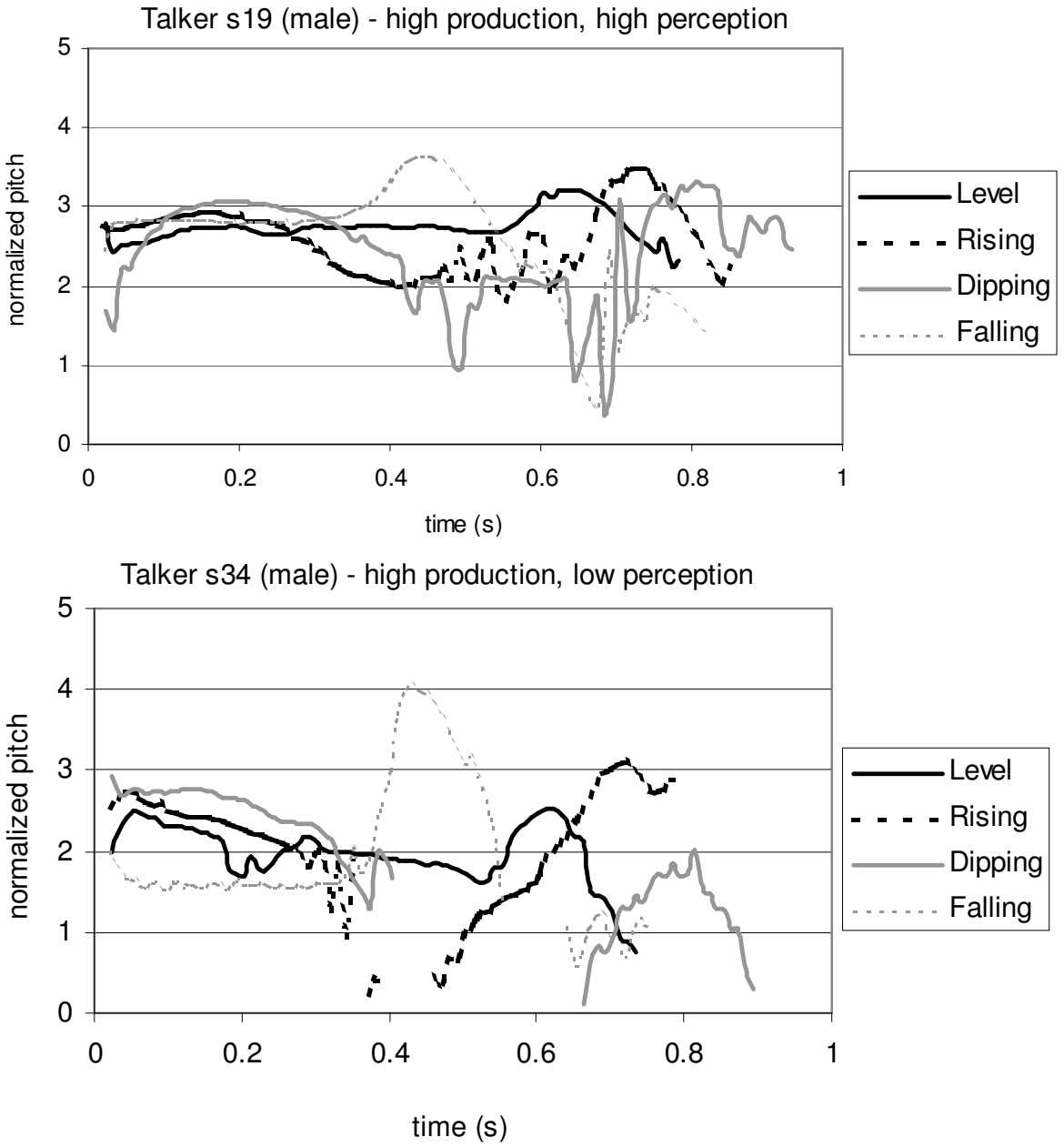


Figure 4.17: Tri-syllabic pitch contours for two talkers, s19 shown on top and s34 shown on bottom, with high production identification scores but very different perception sensitivity scores.

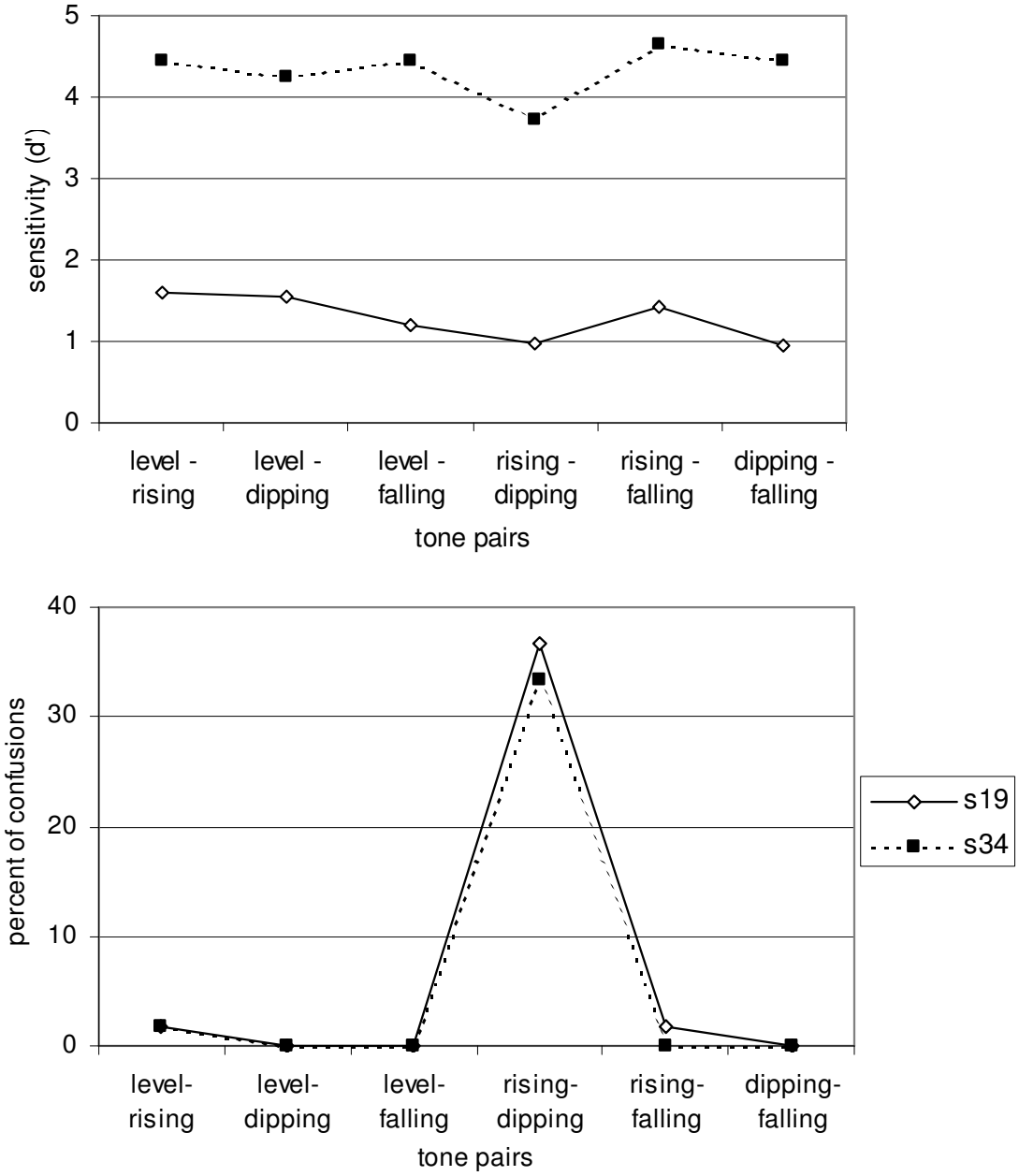


Figure 4.18: Sensitivity scores (shown on top) and production confusion scores (shown on bottom) for two participants, s19 and s34. Sensitivity scores are shown for each tri-syllabic tone pair and tone confusion scores are calculated based on the percent of times that each target tone within the pair was misidentified as the other tone.

These two talker's production abilities are virtually identical with both talkers having difficulty distinguishing the rising and dipping tones but no other tone pairs. However, their perception abilities do not reflect this difference in production accuracy as participant s19 had poorer d' scores across all tone pairs with no clear dip for the rising-dipping tone pair and participant s34 had much higher d' scores with only a slight dip for the rising-dipping pair. This result (for s19) demonstrates an instance in which a non-native's production abilities seem to precede his perception abilities. While these two talkers demonstrated very similar scores from the Mandarin judge's identification scores, their pitch tracks look very different. For example, Talker s34 produced tones with a wide pitch range and more glottalization than Talker s19. How these differences in production may relate to perception abilities needs to be explored further.

Below, in Figure 4.19, are the tri-syllabic pitch contours for two participants, s38 and s50, who performed very similarly and very accurately on the perception task (d' scores of 3.5 and 3.8, respectively), but had widely different scores on the production task (49.2 and 82.5, respectively). In Figure 4.20, these two talkers' sensitivity scores and production confusion scores are shown.

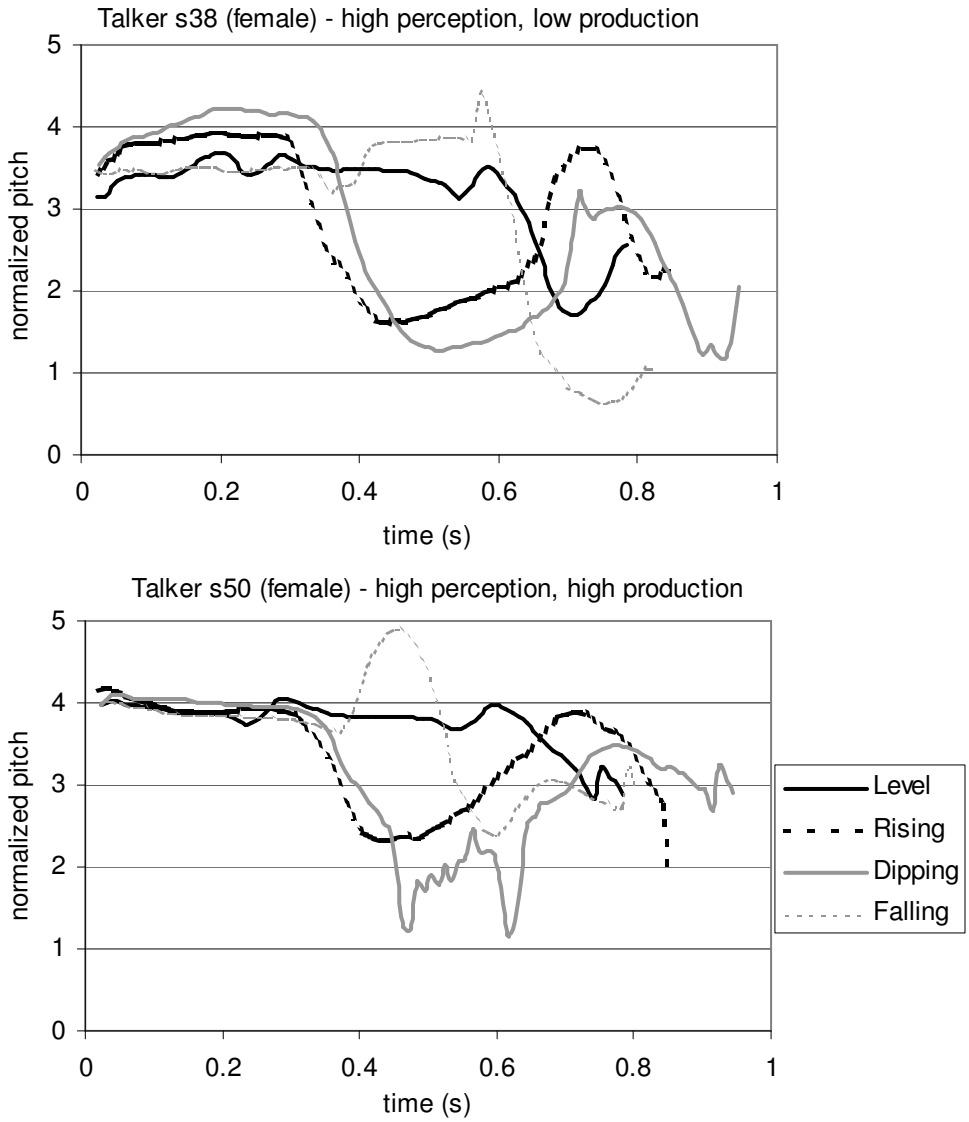


Figure 4.19: Tri-syllabic pitch contours for two talkers with divergent production abilities and similar perception abilities. Talker s38's productions (shown on top) were identified poorly while talker s50's productions (shown on bottom) were identified very accurately.

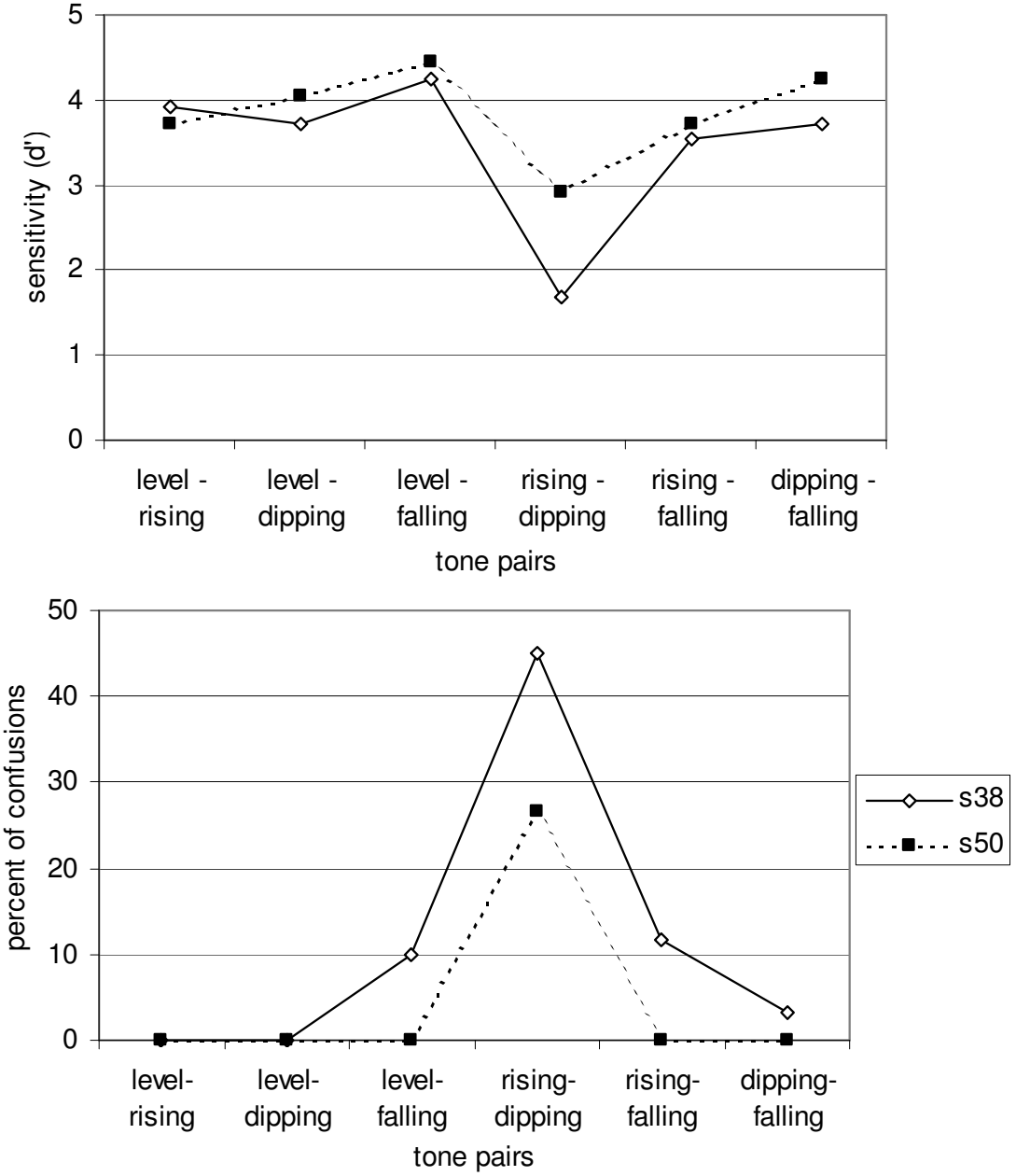


Figure 4.20: Sensitivity scores (shown on top) and production confusion scores (shown on bottom) for two participants, s38 and s50. Sensitivity scores are shown for each tri-syllabic tone pair and tone confusion scores are calculated based on the percent of times that each target tone within the pair was misidentified as the other tone.

For this pair of participants, there seemed to be a stronger relationship between their perception and production abilities on the tri-syllabic stimuli compared to talker pairs discussed above for the monosyllabic and tri-syllabic stimuli. In both perception and production, both participants performed least accurately on the rising-dipping tone pair. Furthermore, participant s38 performed less accurately than participant s50 in both perception and production for this tone contrast. Talker s50's productions of all other tone pairs were never confused with one another while talker s38 had substantially more difficulty producing the falling tone as it was frequently confused with other tones. Talker s38's production of the falling tone was highly deviant from the Mandarin model since it was level during the middle syllable rather than showing a falling pattern. Furthermore, this talker's rising and dipping tones were produced in a nearly identical manner. However, there is not a substantial difference between the two talkers' perceptual sensitivity to the pairs with the falling tone, indicating a discrepancy between perception and production abilities.

In sum, while some evidence was found supporting a connection between perception and production abilities for particular tone pairs for the tri-syllabic stimuli, there were more instances in which these two abilities diverged.

4.4 Discussion

4.4.1 Mandarin Identification Judgments and Acoustic Analysis

Overall, the English listeners' imitations of the Mandarin tones were very accurately identified by the native Mandarin judges for both the monosyllabic and tri-

syllabic stimuli, 72% and 73% respectively. These overall identification accuracy scores were even higher than those reported in Wang et al. (2003) in which native English talkers' monosyllabic productions were identified with 57% accuracy before perceptual training. The participants in Wang et al.'s study had studied Mandarin for one or two semesters. This difference between the current study and Wang et al. can be accounted for by two factors. First, the participants in the Wang et al. study read the stimulus materials rather than imitating a Mandarin model. It is likely more difficult to produce a non-native category from memory than to imitate one. Second, the stimulus materials in Wang et al. included syllables with a variety of consonants and vowels whereas the current study used only the syllable /ra/. This segmental variability may have increased the task complexity and made it more difficult for the talkers to accurately produce the tonal contours. The overall success at imitating the Mandarin tone categories found in the current study suggests that experience with English intonation did not severely hinder the talkers' abilities to produce the Mandarin tones. However, since the participants had just completed the perceptual discrimination task, they may have performed more accurately than native English speakers with absolutely no exposure to Mandarin tones.

The acoustic analyses indicated that the English talkers were able to imitate shape of the Mandarin tones. Nearly all the deviations from the Mandarin model stemmed from deviations in pitch height rather than contour shape. The types of production errors were similar for the monosyllabic and tri-syllabic stimuli and were comparable to errors observed in other studies.

For the level tone, the Mandarin judges were able to very accurately identify the talkers' productions. The deviations for this tone compared to the Mandarin model involved a slightly lower pitch height for the entire contour. This tendency to produce the level tone in a slightly lower pitch range than the Mandarin model may be caused by transfer from English. In English intonation, there are no pitch patterns which maintain a level contour at the top of the pitch range, but there are level contours with a mid-range pitch height.

The rising tone was often confused with the dipping tone particularly in the tri-syllabic condition. While the shape of the rising contour was generally accurate, it started too high and ended too low in the monosyllabic condition. Similarly, in the tri-syllabic condition, the surrounding tones were lower than the Mandarin model resulting in starting and ending points that were too low.

The dipping tone was identified frequently as the falling tone in the monosyllabic condition whereas it was most often confused with the rising tone in the tri-syllabic condition. A larger variety of longer utterances will have to be investigated to determine if these confusion patterns are a result of the particular surrounding tonal context investigated in this study or if different contexts would produce similar results. In both the monosyllabic and tri-syllabic conditions, the dipping tone did not have a low enough valley. Furthermore, in the tri-syllabic context the height of the surrounding tones were too low. If the patterns of perceptual discrimination difficulties lead to difficulties in production (as was found with the tri-syllabic rising-dipping contrast) then the patterns of production accuracy would vary depending on the tones of the surrounding syllables.

Further research is needed to determine if the types of deviations from the Mandarin model differ depending on the surrounding tonal context.

The falling tone was identified very accurately in both the monosyllabic and tri-syllabic conditions. The falling tone in both conditions had a peak that was not as high as the Mandarin model and an overall smaller pitch range.

For the rising, dipping, and falling tones, the overall compressed pitch range including starting and ending points and peaks and valleys that were closer to the middle of the pitch range compared to the Mandarin model may be a result of transfer from English intonation. First, the overall pitch range in English has been found to be smaller than that for Mandarin (Chen, 2005). Second, the rate of pitch change in Mandarin is greater than in English due to the unit size over which the pitch patterns are placed (Eady, 1982). Native English talkers may have difficulty realizing such rapid rises and falls within the domain of one syllable. Therefore, they produce the tones with a smaller pitch range and smaller rises and falls, while still producing an accurate imitation of the shape of the contour.

The ways in which the English talkers in the current study deviated from the Mandarin model are very similar to the pre-perceptual training results from Wang et al. (2003) for monosyllabic stimuli using a reading task to elicit tone production even though the participants in the two studies differed in their exposure to Mandarin (i.e. some experience for the Wang et al. participants and no prior experience for the current participants) The only notable discrepancy between the results in the current study and the Wang et al. results was that in the current study the monosyllabic dipping tone was

most frequently identified as the falling tone whereas in the Wang et al. study it was most frequently identified as the rising tone. This difference could have resulted from the differences in the tasks (reading versus imitation) or from difference in the participants experience with Mandarin (some versus none). Also, the particular talker chosen as the Mandarin model may have lead to this pattern of errors since his dipping tone did not have as extensive a rising portion as some Mandarin talkers produce.

The current study extends previous findings by examining tone production in utterances longer than a monosyllable. The errors were very similar across the monosyllabic and tri-syllabic conditions. The only notable difference between the monosyllabic and tri-syllabic conditions was that the dipping tone was most frequently identified as the falling tone in the monosyllable condition whereas it was most frequently misidentified as the rising tone in the tri-syllabic condition. This finding may suggest that the English listeners are using the same set of categories (presumably English prosodic categories) to guide their imitation of the Mandarin tones regardless of the length of the utterance and therefore their deviations are similar in the two conditions. A more extensive set of contextualized tones will have to be examined in order to more fully characterize the ways in which English prosody affects Mandarin tone production.

4.4.2 Relationship between Perception and Production

There was not a strong link between individual participants' perception and production abilities in this study. This lack of a relationship was observed both when all the participants' perception and production scores were correlated and when the ranking

of tone pair scores for perception and production were compared (either for all participants averaged or for selected individual participants). However, the ranking of tone pairs was much closer to reaching significance for the tri-syllabic stimuli than the monosyllabic stimuli. This lack of a perception-production relationship may indicate that with *naïve* non-natives these two abilities are not linked but with experience, the representations in the two modalities will become more synchronized. The findings from Wang et al. (2003) in which there was a strong tie between perception and production abilities suggests that this link is experience dependent. Most, if not all, studies of the perception-production link for segmental categories have investigated talker-listeners who have had at least some experience with the target language. Therefore, it is unclear whether this finding is caused by the difference in the type of phonological structure investigated (segments versus prosody) or if it is caused by the differences in the participants' exposure to the non-native language (some versus none). Naïve non-natives' segmental perception-production abilities should be studied to determine if there would be a strong link between perception and production as has been shown for participants with prior exposure to the non-native language.

While individuals' overall perception and production scores did not correlate, there was some evidence of perception leading production for the group data although this result is inconclusive. First, the perceptual sensitivity of the participants was high across nearly all tone pairs in both the monosyllabic and tri-syllabic conditions. Production scores were also generally quite accurate although participants did have some difficulty accurately producing some pairs. These findings suggest that while listeners

were perceptually sensitive to nearly all pairs, production at times lagged. Second, the monosyllabic pair for which the participants had most difficulty in production (the dipping-falling pair) showed high perceptual sensitivity. This finding may indicate an instance of perception leading production. The tri-syllabic pair for which the participants had most trouble in perception (the rising-falling pair) was also the pair they had most difficulty accurately producing. This finding may indicate an instance in which the lack of perceptual sensitivity to a contrast hindered production accuracy. There were no tone pairs, at least for data averaged across the participants, in which sensitivity was very low and production was very high.

The study was limited by the method used to elicit the productions, imitation. With completely naïve participants it is difficult, if not impossible, to get them to produce a contrast for which metalinguistic labels are not available. For this reason, an imitation task was chosen. However, this imitation task may have lead to differences in this study compared with others in which the productions were elicited using different methods.

More acoustic analysis of the data should be performed specifically investigating the ways in which the productions are similar or different to the Mandarin model in terms of voice quality, amplitude envelope and duration. These three cues are important, if secondary to pitch, in the categorization of the Mandarin tones. Furthermore, while not systematically explored in the current study, my impressionistic judgment was that some of the participants, who tended to be the least accurate in production, seemed be attempting to use vowel quality instead of pitch differences to differentiate the tones.

Therefore, formant measurements might be useful in determining other ways in which these speakers deviated from the Mandarin model.

4.5 Conclusion

This study demonstrated that talkers are able to fairly accurately imitate unfamiliar non-native prosodic pitch contours. The ways in which these talkers deviated from the native model were similar to previous studies with talkers with experience with the non-native language. However, there was not a strong link between the participant's perception and production abilities. This finding contrasts with previous findings in both studies of segmental and suprasegmental categories in which a correspondence was found for perception and production abilities for participants with some prior exposure to the target language.

While this study extended previous studies of Mandarin tone production and perception by investigating stimuli longer than a monosyllable, it was limited by using only one tonal frame. Other tonal frames should be tested in order to further investigate the relationship between production of non-native prosodic categories and native language categories. Specifically, variability was observed in the perceptual sensitivity to tone pairs in different tonal frames (in Experiment 2, Chapter 3). Testing whether the variability observed in the perception experiment would be mirrored in a production task would provide more information into the link between perception and production for non-native prosodic categories. This follow-up experiment could also provide insight about

how the prosodic categories in the native and non-native languages influence one another.

A question that this study did not explore is the cause of individual variability in perception, production, and the perception-production relationship. The cause of the individual variability should be explicitly investigated in future studies. Furthermore, all analyses reported in this study were normalized for duration. An investigation of the English listeners' abilities to accurately imitate the systematic differences in duration among the tone categories could also reveal an influence of English prosodic categories on Mandarin tone production.

CHAPTER 5

5.1 Summary

The three experiments in this dissertation tested native English and Mandarin participants' abilities to perceive and produce Mandarin lexical tone contrasts in monosyllabic and tri-syllabic utterances. The perception of the tone contrasts was tested in two discrimination experiments in which listeners had to indicate whether the members in lexical tone pairs were the same or different. These responses were used to calculate the listeners' sensitivity to the tone contrasts. Reaction times to these trials were entered into a multidimensional scaling analysis and used to determine the perceived similarity of the tones and the acoustic features to which the listeners were attending. Native English participants' production abilities were tested in an imitation task. The productions were evaluated by native Mandarin listeners and through acoustic analysis. The central findings of these experiments are listed below. Following the description of the main findings is a discussion of the main results with particular emphasis on the implications of the current findings for models of cross-language speech perception. Suggestions for future work are also offered.

1. The native English listeners were overall less sensitive to the Mandarin lexical tone contrasts than the Mandarin listeners.
2. While the native English listeners were less sensitive to the Mandarin lexical tone contrasts than the Mandarin listeners, they displayed overall fairly high perceptual sensitivity to Mandarin tone contrasts both for monosyllabic and tri-syllabic stimuli.

3. Mandarin listeners' perceptual sensitivity to the tone contrasts in the tri-syllabic stimuli was high regardless of the tonal frame in which the target tones were presented (except for one tone pair in one tonal frame). In contrast, the English listeners' perceptual sensitivity to tone pairs varied widely across the tonal frames. The variability in sensitivity was significantly related to the acoustic similarity between the pitch contours to be discriminated.
4. For both the Mandarin and English listeners, the acoustic features attended to varied across the monosyllabic and various tri-syllabic conditions. Where determinable, these perceptual dimensions could be related to features that are important in distinguishing prosodic categories in each native language system. Mandarin listeners mostly attended to lexical tone targets whereas English listeners attended mostly to global aspects of the stimuli.
5. English listeners fairly accurately imitated the Mandarin tonal contours for both the monosyllabic and tri-syllabic stimuli as determined by the Mandarin listener judgments and through acoustic analysis. The deviations from the Mandarin model, mostly involving pitch height and pitch range rather than pitch contour shape, were similar for the monosyllabic and tri-syllabic stimuli.
6. The participants' abilities to accurately perceive and produce the Mandarin tone stimuli were not significantly correlated. That is, participants with high perceptual sensitivity to the tone contrasts were not necessarily highly accurate at tone production and vice versa. Furthermore, individual tone pairs that were most

accurately produced were not necessarily the ones with the highest perceptual sensitivity scores.

5.2 Discussion

There are two central contributions of the work presented in this dissertation. First, the results demonstrated that native and non-native listeners appear to attend to different aspects of the stimuli during lexical tone perception suggesting that the primary mode of processing differs across the two groups (auditory vs. linguistic). Second, the study demonstrated the importance of incorporating contextual variation into the characterization of non-native prosodic perception and production.

5.2.1 Listening modes

The results from the two perception experiments suggest that the English and Mandarin listeners were processing the lexical tones in different primary modes, acoustic for the English listeners and linguistic/phonetic for the Mandarin listeners. The English listeners were processing the tones primarily in the acoustic mode, but they also may have been influenced by linguistic knowledge. The finding that the English listeners' sensitivity scores were significantly correlated with the acoustic similarity of the pairs suggests that acoustic processing was primary. However, if the English listeners were performing in a completely non-linguistic mode, they should have displayed equivalent sensitivity to these contrasts compared to the Mandarin listeners as has been shown with non-tone language listeners' perception of lexical tones when they are low-pass filtered

or played as music (Burnham et al., 1996). Therefore, it seems likely that the lexical tones were not directly assimilated to native language prosodic categories but that linguistic processing generally had some influence on their sensitivity to the contrasts. The lexical tone pairs may have been perceived as "uncategorizable" pairs. Within this type of assimilation pattern the acoustic similarity between the categories as well as their proximity to native categories influences their perception. This type of assimilation pattern most closely fits the current data as there was evidence that the English listeners were significantly influenced by acoustic similarity between the contrasts and that the contours with lower sensitivity scores seemed to approximate assimilation patterns in which lower sensitivity is predicted (in the Perceptual Assimilation Model).

In contrast to the primacy of acoustic processing for the English listeners, the Mandarin listeners appeared to be processing the tones mainly using linguistic processing. The main evidence for this finding was that the Mandarin listeners were virtually insensitive to the acoustic similarity between the members of a pair and were highly sensitivity to almost all pairs. However, there did appear to be a slight influence of acoustic similarity since the one contrast that the Mandarin listeners were significantly less sensitive to (the rising-dipping pair in the level-falling tonal frame) was a pair with one of the high acoustic similarity scores. Therefore, while generally Mandarin listeners process the lexical tones in a linguistic mode and are not significantly influenced by acoustic similarity there may be some cases in which the high degree of acoustic similarity between the contours leads to a lower degree of sensitivity to the contrast.

While significantly less sensitive to the contrasts than the Mandarin listeners, the English listeners showed overall quite high sensitivity to the Mandarin tone contrasts as their d' scores ranged from 2.1 to 4.0 while the Mandarin listeners scores ranged from 2.7 to 4.3. This finding suggests that while experience with English prosodic categories did not prepare the listeners to perfectly discriminate these tone contrasts, their native language experience did not lead to an absolute insensitivity to the contrasts either. While Mandarin and English have pitch patterns that are broadly comparable, the differences across the languages in the details of these patterns clearly affected the native English listeners' overall sensitivity. The English listeners' primary reliance on the acoustic similarity of the contrasts may have allowed them to discriminate these stimuli well.

A remaining question is how listeners from other languages will utilize different processing strategies. Previous research has shown that listeners from a native language that is prosodically similar to the target non-native language will be more sensitive to the contrasts from the non-native language. For example, tone language listeners seem to be afforded an advantage over lexical stress language listeners when perceiving non-native tone contrasts (Gandour et al., 2000; Lee et al., 1996; Francis et al., 2004; however, see Burnham et al., 1996). This advantage could stem from the similarities between the languages in their systems of lexical prosody. For example, Mandarin listeners may be afforded a greater advantage in perceiving Cantonese compared to Yoruba because Mandarin and Cantonese are both contour tone languages while Yoruba only has level tones. However, Mandarin listeners should better than English listeners at perceiving both Cantonese and Yoruba since Mandarin is a tone language and English is not.

A question for future research is exactly why the listeners from prosodically related languages are afforded an advantage in the perception of non-native prosodic contrasts. Listeners from a native language similar to the non-native language might be expected to rely more on linguistic processing rather than auditory processing, which could help or hinder the listeners depending on the correspondence between the categories in the native and non-native languages. An analysis of assimilation patterns and reliance on auditory cues (as in the current study) would help to reveal how these two modes of listening influence listeners from various languages. Listeners may shift their reliance on different types of knowledge (i.e. linguistic versus acoustic) depending on the task and the relationship between the contrasts presented and their native system of contrasts. While in segmental perception studies it appears that non-native listeners generally process segments linguistically (except for discrimination experiments with very short ISIs), prosodic perception may differ in that listeners are able to at least partially disregard linguistic knowledge and primarily focus on the acoustic similarity of the categories. However, it should be noted that the task conditions used in the current experiments (relatively short ISI, same-different task) may have also encouraged or allowed an acoustic processing strategy which listeners may not be able to utilize in other types of experimental paradigms.

The behavioral results in this study are in accord with neurological studies in which listeners from non-tone languages have been shown to use different parts of their brains compared to tone language listeners when processing lexical tone stimuli (e.g. Wang, Jongman, and Sereno, 2001; Wong et al., 2004). The recruitment of different

cortical areas mirrors the behavior results here in which the Mandarin and English listeners appear to be utilizing different processing modes.

5.2.2 Importance of context

The results from this experiment showed that English listeners' sensitivity to tone pairs differed depending on whether Mandarin lexical tones were presented in isolation or in tri-syllabic utterances and also differ depending on the tones of the surrounding syllables. This finding extends previous research in which sensitivity was only measured with monosyllabic stimuli by demonstrating that English listeners are less sensitive than native Mandarin listeners with longer utterances as well. Furthermore, English listeners were less sensitive overall compared to the Mandarin listeners for both the isolated monosyllables and the tri-syllabic stimuli. The lower sensitivity for the English listeners is not merely due to their unfamiliarity with the pitch patterns applied over monosyllables, as shown by the tri-syllabic results, but most likely involves differences between Mandarin and English in terms of their prosodic typology and in the way these categories are phonetically implemented with regard to the details of timing, alignment, and relative pitch height and range.

The Mandarin listeners' patterns of sensitivity to the tone pairs were relatively consistent regardless of the tonal frame in which the target tones were presented. This finding suggests that native listeners are typically able to extract the underlying pitch targets in coarticulated tones and consequently correctly discriminate the tone pairs. In contrast, the native English listeners' sensitivity to the tone pairs varied depending on the

tonal frame in which they were presented. The variability in sensitivity probably resulted from the variability in the acoustic similarity between coarticulated tone pairs and may have been influenced by differences in the ways the coarticulated tones are related English prosodic categories as well. These results parallel findings for non-native segmental perception as previous studies have shown that the similarities among native and non-native segmental categories shift depending on their realization in context. This contextual variation influences the accuracy with which the non-native listeners are able to identify the non-native segments (e.g. Logan, Lively, and Pisoni, 1991; Strange et al., 2004). The current results and results from segmental studies demonstrate the importance of investigating cross-language speech perception in utterances longer than a syllable or word. A comprehensive model of cross-language speech perception will need to explain the interactions of native and non-native categories within the segmental and suprasegmental domains both in their isolated forms and their multiple contextualized versions.

The multidimensional scaling results revealed an influence of context for the Mandarin listeners when comparing scaling solutions for the monosyllabic versus tri-syllabic stimuli in Experiment 1 (Chapter 2). While rising and dipping tones were in close proximity for the monosyllabic solution (replicating Huang, 2004), they were well separated in the tri-syllabic solution. This result may be a case in which perceptual distinctiveness is enhanced when stimuli are put into longer utterances. While isolated forms of tones may be considered canonical, the information presented in the preceding and following tones from anticipatory and carry-over articulation may provide more

information about the tone's identity than when the tone is presented in isolation. This finding may not be surprising since real world speech communication depends on listeners abilities to extract underlying targets from running speech.

Contextual variation also influenced the production results. While most of the types of acoustic deviations for the English productions compared to the Mandarin model were similar across the monosyllabic and tri-syllabic conditions, there was one case in which the results for the monosyllabic and tri-syllabic conditions differed, the dipping tone. In the monosyllabic productions the dipping tone was most frequently misidentified as falling whereas in the tri-syllabic production it was most often misidentified as rising. This discrepancy probably occurred due to the differences in the production of the dipping tone in the tri-syllabic frame compared to the tone's production in isolation. In the tri-syllabic frame the pitch at the end of the dipping tone had to rise in anticipation of the following level tone. This pattern of tonal coarticulation most likely resulted in the confusion with the rising tone. The dipping tone in isolation in contrast seemed to lack a rising portion completely and therefore was confused with the falling tone. As with the perception data, these types of discrepancies point to the importance of including longer stimulus items when testing non-native speech production. While these findings are suggestive, more tri-syllabic contexts should be investigated in order to determine the cause of these differences across monosyllabic and tri-syllabic conditions.

5.3 Implications for models of cross-language speech perception

The current results suggest several areas in which models of cross-language speech perception need to be extended in order to account for non-native prosody perception. The four central expansions of the models suggested by the work in this dissertation are discussed below.

First, a comprehensive model of cross-language speech perception needs to account for the different types of information that non-native listeners can recruit during speech perception (i.e. listening modes). The Perceptual Assimilation Model is the only model which is explicit about listeners' perception of speech in a speech versus non-speech mode. For prosody perception the "uncategorizable" type of assimilation pattern may need to be further specified as it appears that in the current experiment, listeners are using both linguistic and auditory processing when discrimination the non-native stimuli. The PAM makes predictions regarding sensitivity based on "similarity between the categories" but the way(s) in which similarity should be calculated for prosodic categories needs to be determined in order to make explicit predictions. A comprehensive model would also need to have an account for how listeners use information from the two perception modes and how the processing shifts from a primarily auditory mode to a presumably primarily linguistic mode during the development of non-native speech perception. Hypotheses about how the similarities between the native and non-native languages will effect listening mode will also have to be developed. While some of the ideas presented in the current models of cross-language speech perception can provide insight into some of the factors which might influence

non-native prosodic perception, it appears that these models will have to undergo extensive revision to fully account for the perception of all levels of non-native phonological structure.

Second, models of cross-language speech perception have to account for contextual variation. While the Speech Learning Model considers position sensitive allophones to be the unit of analysis, the Perceptual Assimilation Model and the Native Language Magnet Model both focus their hypotheses on phonemes. For predictions regarding how listeners from one language will perceive and produce the categories from another language to be accurate, these models need to take into account how contextual variation, including coarticulation, affects discrimination, identification, and production. This dissertation has shown how contextual variability for prosodic categories changes acoustic similarity between categories and discrimination patterns between native and non-native categories. Since patterns of coarticulation are language specific (e.g. Manuel and Krakow, 1984), this factor needs to be considered when considering the mapping from one system to another. Incorporating contextual variation is vitally important because real world speech communication typically involves running speech, as few speech utterances are composed of isolated monosyllables. Therefore, the current research represents a small step towards the characterization of naturalistic speech perception and production.

Third, these results demonstrated that English listeners' difficulties with the Mandarin tone contrasts are not simply a problem in categorizing the stimuli into linguistically relevant categories or a problem of lexical access. The use of a

discrimination task is essential in separating the decreases in sensitivity from lack of metalinguistic knowledge. The extraction of meaning or the labeling of categories will present non-native listeners with additional sources of native language interference. The overlap in the two listener groups' sensitivity scores in the current study contrasts with an investigation of *identification* in which no overlap between native and non-native listener groups was observed (Bent, Bradlow and Wright, in press). This task dependent difference suggests that models of cross-language speech perception need to include both sensitivity- and categorization-related effects.

Fourth, while studies of non-native segmental perception have shown both loss of sensitivity to non-native segmental contrasts due to lack of exposure and enhanced distinctiveness of segmental contrasts due to native language exposure, the current study has shown maintenance of sensitivity to non-native prosodic contrasts. The learning trajectory for prosodic categories needs further study in order to determine whether the maintenance of sensitivity to non-native prosodic categories is consistent across native/non-native language pairs and across different dimensions of prosodic structure.

In sum, models of cross-language speech perception will have to be substantially revised to account for the perception and production of non-native prosody. The work presented in this dissertation provides some tentative support for a model in which sensitivity to prosodic contrasts is maintained because listeners are able to recruit acoustic processing during non-native prosody perception. Furthermore, the contribution of contextual variation to non-native listeners' sensitivity has been shown to be

substantial suggesting that in a comprehensive model of cross-language speech perception the unit of analysis must be more detailed than isolated prosodic categories.

5.4 Future research

Current frameworks (e.g. Jun, 2005) in which languages are described by their lexical prosody, intonation, rhythmic structure, and prosodic unit size(s) will be useful in determining the ways in which the structure of any native language will influence the perception and production of any non-native language. While the work in this dissertation focused on intonation and lexical prosody, all aspects of suprasegmental structure can vary across languages and are important to investigate in order to fully characterize cross-language speech perception. While listeners from a language which shares certain features with the non-native language may be afforded an advantage in speech perception and production, it remains to be determined which specific features across languages influence sensitivity to non-native prosodic contrasts. The investigation of longer utterances is essential in determining how larger prosodic units interact during non-native speech perception.

Much more work investigating the sensitivity of listeners from a wide variety of language backgrounds needs to be conducted to determine what type of prosodic categories influence listeners' perception of non-native speech. While some work has suggested interference from a native lexical stress language during the *production* of a non-native lexical stress language in terms overall intonation and pitch accent patterns (e.g. Buysschaert, 1990; Grover, Jamieson, and Dobrovolsky, 1987; Mennen, 2004), the

expansion of this work needs to compare both perceptual sensitivity to and production of prosodic categories across languages of different prosodic types (e.g. Guion, to appear). While the current investigation included the study of the perception and production of Mandarin lexical tones by English listeners, a clear follow-up study would be the investigation of the perception and production of the English intonation by speakers of Mandarin or another lexical tone language. It would be particularly illuminating to determine whether the lexical tone listeners would rely primarily on auditory information during the perception of English intonation as the English listeners in this study did during the perception of Mandarin lexical tones.

Future investigations should include meaningful utterances in addition to non-word ones. Studies should be conducted to determine how different types of information within the target language interact with one another during cross-language speech perception and production and further how the types of meaning associated with the prosodic categories in the native and non-native languages affects perception and production. Conducting studies of this type will require participants with experience with the non-native language. While most studies of Mandarin tone perception by English listeners have focused on listeners with at most two years of classroom study (Gottfried and Suiter, 1997 is one exception in which advanced listeners were included), most studies of the interaction of intonation systems have included highly proficient non-natives. The testing of different language combinations also needs to include participants with varying degrees of exposure to the target language.

Lastly, further investigation of the two central factors identified in this dissertation as important for characterizing cross-language speech perception, listening mode and contextual variation, need to be investigated for a wider range of languages. The languages should include those which vary along the various dimensions on which prosodic systems differ.

5.5 Summary

The experiments in this dissertation examined the perception and production of Mandarin lexical tones by native English speakers. The stimuli used in the experiments were both monosyllables and tri-syllabic utterances. The primary objective of this research was to determine how models of cross-language speech perception would need to be expanded or revised to account for non-native perception and production of suprasegmental phonological structure. The English speaking participants' high sensitivity to the Mandarin lexical tones and their overall success at imitating the tones contrasts with most findings in non-native segmental perception and production. However, the English participants showed more variable sensitivity to the Mandarin tones and appeared to be relying more on acoustic similarity than the Mandarin participants. These differences suggest that the English listeners were processing the tones in a primarily acoustic mode while the Mandarin listeners processed the tones linguistically. Furthermore, the inclusion of multiple longer stimuli demonstrated the importance of incorporating contextual effects when characterizing non-native prosodic perception. A comprehensive model of cross-language speech perception will need to

account for how native and non-native listeners utilize different modes of listening and must incorporate contextual variation. While some of the factors identified in current models of cross-language speech perception are also important for prosodic perception, the models will need to be substantially expanded to account for the non-native perception and production of all levels of phonological structure.

REFERENCES

Anderson-Hsieh, J., Johnson, R. & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure, *Language Learning,* 42(4), 529-555.

Beckman, M, E. & Elam, G. A.. (1997). Guidelines for ToBI labelling, version 3.0. Manuscript and accompanying speech materials, Ohio State University.

Beckman, M. & Pierrehumbert, J. (1986). Intonational structure in Japanese and English, *Phonology Yearbook III*, 15-70.

Best, C. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 171-204). Baltimore: York.

Best, C. T., Faber, A., & Levitt, A. G. (1996). Perceptual assimilation of non-native vowel contrasts to the American English vowel system. *Journal of the Acoustical Society of America*, 99, 2602.

Best, C., McRoberts, G., & Sithole, N. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 45-60.

Best, C., McRoberts, G., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system, *Journal of the Acoustical Society of America,* 109, 775-794.

Best, C. & Strange, W. (1992). Effects of phonological and phonetics factors on cross-language speech perception on approximants, *Journal of phonetics*, 20, 305-330.

Bent, T., Bradlow, A. R. & Wright, B. (in press). The influence of linguistic experience on the cognitive processing of pitch in speech and non-speech sounds. *Journal of Experimental Psychology: Human Perception and Performance.*

Blitcher, D., Diehl, R. & Cohen, L. (1990). Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: evidence of auditory enhancement, *Journal of Phonetics*, 18, 37-49.

Bluhme, H. & Burr, R. (1972). An audio-visual display of pitch for teaching Chinese tones. *Studies in Linguistics*, 22, 51-57.

Bot, K. de (1980). Relative reliability in judging intonation. *International Journal of Psycholinguistics*, 19, 81-92.

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

Broselow, E., Hurtig, R. R., & Ringer, C. (1987). The perception of second language Prosody. In G. Ioup & S. H. Weinberger (Eds.), *Inter-language Phonology, The Acquisition of a Second Language Sound System* (pp. 350-361). Cambridge: Newbury House.

Burnham, D., Francis, E., Webster, D., Luksaneeyanawin, S., Attapaiboon, C., Lacerda, F., & Keller, P. (1996). Perception of lexical tone across languages: Evidence for a linguistic mode of processing. In T. Bunnell & W. Idsardi (Eds.), *Proceedings of the Fourth International Conference on Spoken Language Processing. Vol I*, 2514-2517.

Burnham, D. & Mattock, K. (to appear). The perception of tones and phones. In Munro, M. & Bohn, O.-S. (Eds.), *Festschrift for James E. Flege*.

Buysschaert, J. (1990). Learning intonation. In J. Leather & A. James (Eds.), New sounds 1990 (pp.300-306). Amsterdam.

Chan, S. W., Chuang, C-K., & Wang, W. S-Y. (1975). Cross-linguistic study of categorical perception for lexical tone. *Journal of the Acoustical Society of America*, 58, S119.

Chao, Y. R. (1948). *Mandarin primer*. Cambridge: Harvard University Press.

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

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

Chen, G. T. (1972). *A comparative study of pitch range of native speakers of Midwestern English and Mandarin Chinese: An acoustic study*. Unpublished doctoral dissertation, University of Wisconsin, Madison

Chen, S. H. (2005). The effects of tones on speaking frequency and intensity ranges in Mandarin and Min dialects. *Journal of the Acoustical Society of America*, 117, 3225-3230.

Chen, Q. (1997). Toward a sequential approach for tonal error analysis. *Journal of the Chinese Language Teachers Association,* 32 (1), 21-39.

Chen, Q. (2001). *Analysis of Mandarin tonal errors in connected speech by English speaker American adult learners: A study at and above the word level*. Unpublished doctoral dissertation, Brigham Young University, Utah.

Chiang, T. (1979). Some interferences of English intonation with Chinese tones. *IRAL,* 17 (3), 245-250.

Chuang, C. K., Hiki, S., Sone, T., & Nimura, T. (1972). The acoustical features and perceptual cues of the four tones of standard colloquial Chinese. In *Proceedings of the 7th International Congress of Acoustics* (Vol. 3). Budapest: Akademial Kiado. pp. 297-300.

Chun, D. (1982). *A contrastive study of the suprasegmental pitch in modern German, American English and Mandarin Chinese*. Unpublished doctoral dissertation, University of California, Berkeley.

Curtis, D. W., Paulos, M. W., & Rule, S. J. (1973). Relation between disjunctive reaction-time and stimulus differences. *Journal of Experimental Psychology*, 99 (2), 167-173.

Derwing, T. & Munro, M. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition,* 19 (1), 1-16.

Eady, S. J. (1982). Differences in the f0 patterns of speech: Tone language versus stress language. *Language and Speech*, 25, 29-42.

Flege, J. E. (1986). The production and perception of foreign language speech sounds. In H. Winitz (Ed.), *Human Communication and its Disorders* Vol. 2 (pp. 224-401). Norwood, NJ: Ablex.

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

Flege, J.E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America,* 84, 70-79.

Flege, J. E. (1993). Production and perception of a novel, second language phonetic contrast. *Journal of the Acoustical Society of America*, 93, 1589-1608.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 233-277). Baltimore: York.

Flege, J. E. (1999). Age of learning and second-language speech. In D. P. Birdsong (Ed.), *New Perspectives on the Critical Period Hypothesis for Second Language Acquisition.* Hillsdale, NJ: Erlbaum.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25 (4), 437-470.

Flege, J. E. & Eefting, W. (1987). The production and perception of English stops by Spanish speakers of English. *Journal of Phonetics*, 15, 67-83.

Flege, J. E. & MacKay, I. R.A. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition*, 26, 1-34.

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

Flege, J.E., Munro, M. & MacKay, I. (1995a). The effect of age of second language learning on the production of English consonants. *Speech Communication,* 16, 1-26.

Flege, J.E., Munro, M. & MacKay, I. (1995b). Factors effecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, 97, 3125-3134.

Flege, J. E., Schirru, C., & MacKay, I. R. A. (2003). Interaction between native and second language phonetics subsystems. *Speech Communication*, 40, 467-491.

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

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics,* 14, 3-28.

Fox, R. A., Flege, J. E., & Munro, M. J. (1995). The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis. *Journal of the Acoustical Society of America*, 97, 2540-2551.

Francis, A., Ciocca, V. & Ma, L. (2004). Effects of native language experience on perceptual learning of Cantonese lexical tones. *Journal of the Acoustical Society of America*, 115, 2544.

Fu, Q-J., Zeng, F-G., Shannon, R. V., & Soli, S. D. (1998). Importance of tonal envelope cues in Chinese speech recognition. *Journal of the Acoustical Society of America*, 104, 505-510.

Gandour, J. & Harshman, R. (1978). Crosslanguage differences in tone perception: A multidimensional scaling investigation. *Language and Speech,* 21, 1-33.

Gandour, J., Wong, D., & Hutchins, G. (1998). Pitch processing in the human brain is influenced by language experience. *NeuroReport*, 9, 2115-2119.

Gandour, J., Wong, D., Hsieh, L., Weinsapfel, B., Van Lancker, D., & Hutchins, G. (2000). A cross-linguistic PET study of tone perception. *Journal of Cognitive Neuroscience*, 12 (1), 207-222.

Gottfried, T. L. & Suiter, T. L. (1997). Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. *Journal of Phonetics*, 25, 207-231.

Grabe, E., Burton, S. R., Garcia-Albea, J. E., & Zhou, X. (2003). Perception of English intonation by English, Spanish, and Chinese listeners. *Language and Speech*, 46, 375-401.

Grieser, D. & Kuhl, P. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology,* 25, 577-588.

Grover, C., Jamieson, D. G., & Dobrovolsky, M. B. (1987). Intonation in English, French and German: Perception and production. *Language and Speech*, 30 (3), 277-296.

Guion, S. G., Flege., J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America*, 107, 2711-2724.

Guion, S.G. (to appear). Knowledge of English word stress patterns in early and late Korean- English bilinguals. *Studies in Second Language Acquisition*.

Guion, S.G., Harada, T. & Clark, J.J. (2004). Early and late Spanish-English bilinguals' acquisition of English word stress patterns. *Bilingualism: Language and Cognition,* 7, 207-226.

Halle, P. A., Best, C. T., & Levitt, A. (1999). Phonetic versus phonological influences of French listeners' perception of American English approximants. *Journal of Phonetics,* 27, 281-306.

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

Hombert, J. M. (1976). Perception of tones of bisyllabic nouns in Yoruba. *Studies in African Linguistics, Supplement 6*, 109-121.

Huang, T. (2001) The interplay of perception and phonology in tone 3 sandi in Chinese Putonghua. In E. Hume & K. Johnson (Eds.), Studies on the Interplay of Speech Perception and Phonology, (pp. 23-42). Ohio State University Working Papers in Linguistics, 55.

Huang, T. (2004). Language-specificity in auditory perception of Chinese tones. Unpublished Doctoral Dissertation, Ohio State University, Columbus.

Ioup, G. & Tensomboon, A. (1987). The acquisition of tone: A maturational perspective. In R. C. Scarcella & M. H. Long (Eds.) *Interlanguage Phonology: Issues in Second Language Research series* (pp. 333-349). Cambridge: Newbury House Publishers.

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

Iverson, P. & Kuhl, P. (1996). Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/. *Journal of the Acoustical Society of America,* 99 (2), 1130-1140.

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.

Jilka, M. (1999). Intonational foreign accent in the speech of American speakers of German. *Journal of the Acoustical Society of America*, 105, 1094.

Jun, S-A. (2005). *Prosodic Typology: The phonology of intonation and phrasing*. Oxford: Oxford University Press.

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

Kewley-Port, D. & Atal, B. (1989). Perceptual differences between vowels located in a limited phonetic space. *Journal of the Acoustical Society of America,* 85 (4), 1726-1740.

Kiriloff, C. (1969). On the auditory perception of tones in Mandarin. *Phonetica,* 20, 63-67.

Kratochvil, P. (1987). The case of the third tone. In Ma (Ed.), *Wang Li memorial volumes* (pp. 253-276). Hong Kong: The Chinese Language Society of Hong Kong.

Kratochvil, P. (1998). Intonation in Beijing Chinese. In D. Hirst & A. DiCristo (Eds.), *Intonation Systems: A Survey of Twenty Languages* (pp. 417-431). Cambridge: Cambridge University Press.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics,* 50, 93-107.

Kuhl, P., & Iverson, P. (1995). Linguistic experience and the perceptual magnet effect, In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 121-154). Baltimore: York.

Kuhl, P. & Meltzoff, A. (1996). Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America*, 100, 2425-2438.

Ladd, D. R. (1996). *Intonational phonology*. New York: Cambridge University Press.

Ladd, D. R., Silverman, K., Tolkmitt, F., Bergmann, G., & Scherer, K. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *Journal of the Acoustical Society of America*, 78, 435-444.

Ladd, D. R. & Morton, R. (1997). The perception of intonation emphasis: continuous or categorical? *Journal of Phonetics*, 25, 313-342.

Ladd, D. R. & Schpman, A. (2003). "Sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of Phonetics*, 31 (1), 81-112.

Leather, J. (1990). Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In J. Leather & A. James (Eds.), *New Sounds 90* (pp.72-97). University of Amsterdam.

Lee, Y-S., Vakoch, D. A. & Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research*, 25, 527-542.

Levey, S. (2004). Discrimination and production of English vowels by bilingual speakers of Spanish and English. *Perceptual and Motor Skills,* 99 (2), 445-462.

Liberman, A. M. & Mattingly, I. G. (1989). A specialization for speech perception. *Science,* 245, 489-494.

Logan, J. D., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, 89, 874-886.

Long, M. (1990). Maturational constraints on language development. *Studies in Second Language Acquisition,* 12, 251-285

MacKay, I. R. A., Flege, J. E., Piska, T., & Schirru, C. (2001). Category restructuring during second-language speech acquisition. *Journal of the Acoustical Society of America*, 110, 516-528.

Manuel, S. Y. & Krakow, R. A. (1984). Universal and Language Particular Aspects of Vowel-to-Vowel Coarticulation. *Haskins Laboratories Status Report on Speech Research*, pp. 69-78.

Mattock, K. & Burnham, D. (2003). Infants' discrimination of tone in a non-speech context. *Australian Journal of Psychology*, 55, 85, Supplement.

Mennen, I. (2004). Bi-directional interference in the intonation of Dutch speakers of Greek. *Journal of Phonetics*, 32, 543-563.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition,* 29, 143-178.

Miracle, W. C. (1989). Tone production of American students of Chinese: A preliminary acoustic study. *Journal of the Chinese Language Teachers Association*, 24, 49-65.

Moore, C. & Jongman, A. (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America,* 102 (3), 1864-1877.

Newman, R. (1996). Individual differences and the perception-production link. *Journal* of the Acoustical Society of America, 99, 2592.

Peng, S.-H. (1996). Phonetics implementation and perception of place *coarticulation and tone sandhi*. Unpublished doctoral dissertation, Ohio State University, Columbus.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation.* Unpublished Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge.

Pierrehumbert, J. & Hirshberg, J. (1990). The meaning of intonation contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in Communication* (pp. 271-311). Cambridge: MIT Press.

Pike, K. (1948). *Tone Languages*. Ann Arbor: University of Michigan Press.

Polka, L. (1991). Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions, *Journal of the Acoustical Society of America*, 89, 2961-2977.

Polka, L. (1992). Characterizing the influence of native experience on adult speech perception. *Perception and Psychophysics*, 52, 37-52.

Rochet, B. L. (1995). Perception and production of second-language speech sounds by adults. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp.379-410). Baltimore: York.

Rose, P. (1987). Considerations in the normalization of the fundamental frequency of linguistic tone. *Speech Communication*, 6, 343-351.

Ruhlen, M. (1976). *A Guide to the Language of the World*. Language Universals Project, Stanford University.

Schack, K. (2000). Comparison of intonation patterns in Mandarin and English for a particular speaker. In K. M. Crosswhite & J. McDonough (Eds.), *University of Rochester Working Papers in the Language Sciences* (pp. 24-55).

Schmidt, A. M. & Flege, J. E. (1995). Effects of speaking rate changes on native and non-native production. *Phonetica*, 52, 41-54.

Scuffil, M. (1982). *Experiments in Comparative Intonation—A Case Study of English and German*. Tubingen: Max Niemeyer.

Shen, X. (1989). Toward a register approach in teaching Mandarin tones. *Journal of the Chinese Language Teachers Association,* 24(3), 27-47.

Shen, X. (1990) *The Prosody of Mandarin Chinese*. Berkeley: University of California Press.

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

Shepard, R. N., Kilpatrick, D. W., Cunningham, J. P. (1975). Internal representation of numbers. *Cognitive Psychology*, 7 (1), 82-138.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., & Nishi, K. (2001). Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *Journal of the Acoustical society of America,* 109 (4), 1691-1704.

Strange, W., Bohn, O.S., Trent, S.A., & Nishi, K. (2004). Acoustic and perceptual similarity of north German and American English vowels. *Journal of the Acoustical Society of America*, 115, 1791-1807.

Wang, W.S.-Y. (1967). Phonological features of tone. *International Journal of American Linguistics*, 33, 93-105.

Wang, W. S-Y. & Li, K-P. (1967). Tone 3 in Pekinese. *Journal of Speech and Hearing Research*, 10, 629-636.

Wang, Y., Jongman, A., & Sereno, J. A. (2001). Dichotic perception of Mandarin tones by Chinese and American listeners. *Brain and Language*, 78, 332-348.

Wang, Y., Jongman, A. & Sereno, J. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America,*113 (2), 1033-1043.

Wang, Y., Spence, M., Jongman, A. & Sereno, J. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America,* 106 (6), 3649-3658.

Weker, J. F. & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception and Psychophysics*, 37, 1.

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

White, C. (1981). Tonal pronunciation errors and interference from English intonation. *Journal of the Chinese Language Teachers Association,* 16 (2), 27-56.

Willems, N. (1982). *English intonation from a Dutch point of view*. Dordrecht: Foris Publications.

Wong, P. C. M., Parsons, L. M., Martinez, M., & Diehl, R. L. (2004). The role of the insular cortex in pitch pattern perception: The effect of linguistic contexts. *The Journal of Neuroscience*, 24, 9153-9160.

Young, F. W. (1987). *Multidimensional scaling: History, Theory, and Applications*. Hillsdale, NJ: Erlbaum.

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

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

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

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

Xu, Y. & Xu, C. X. (to appear). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics*.

APPENDICES

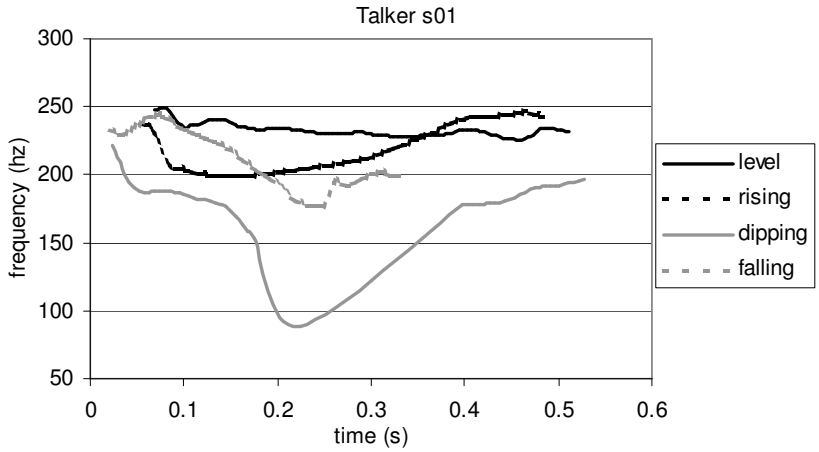
Appendix A: Stimuli for Pilot Study Comparing Mandarin and English Pitch Contours

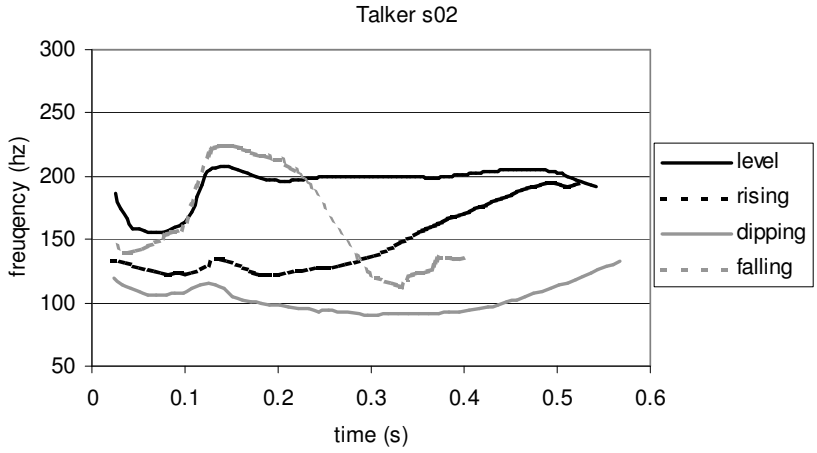
1. ang1 an1
2. wen1 yu2
3. ang1 lian3
4. wan1 lu4
5. ming2 yan1
6. lan2 men2
7. ning2 niao3
8. mian2 mao4
9. min3 ou1
10. lao3 lv2
11. mei3 ma3
12. leng3 maio4
13. le4 mao1
14. li4 mao2
15. le4 lian3
16. li4 yu4
17. mao1ai4 niao3
18. ma1 laio1 la4
19. mao1 ai4 niao3
20. mao1 lan3 yu2
21. ou1 yao4 er2
22. ma1 wei1 ou1
23. mei4 lan3 er2
24. ma1 wei1 ou1
25. ma1 ao2 ou1
26. ni3 ai4 mao1
27. wo3 wei4 yu2
28. wo3 ao2 yu2
29. nan2 wei1 yu2
30. nan2 ao2 yu2
31. ni3 liang2 an1
32. wo3 wei1 ou1
33. nan2 wei1 ou1
34. non2 liang2 an1
35. wo3 wei1 niao3
36. wo3 liang2 ma3
37. ni3 ai2 ma3
38. ni3 liao1 mao4
39. ni3 laing2 liao4
40. nan2 wei1 miao3
41. nan2 mai4 ma

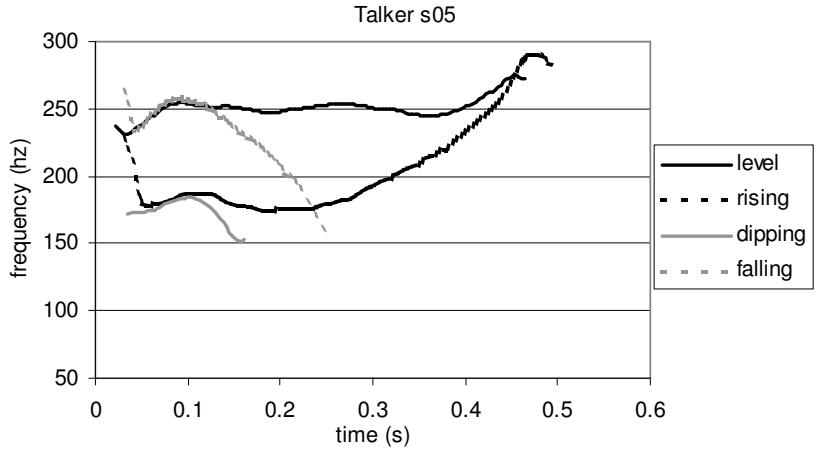
42. wo3 wei1 niao3
43. lan3 er2 wei1 li4 niao3
44. ni3 ai4 ma3
45. nan2 wei1 niao3
46. nan2 mai4 ma3
47. nan2 liao1 mao4
48. lv2 lan3 an1
49. nan2 yao4 mao1
50. er2 mai3 lv2
51. er2 mai4 lv2
52. yu4 ai4 yu3
53. yan4 ai4 mao4
54. mao1 ai4 mao4
55. yu4 mai4 yu2
56. niu1 lan3 wa1
57. ma1 ai4 mao1
58. ma1 wei1 yu2
59. ma1 ao2 yu2
60. niu1 liang2 ma3
61. niu1 liang2 lu4
62. wa1 lan3 yue4
63. nan2 ao2 niao3
64. nan2 liang2 lu4
65. er2 mai3 mao4
66. nan2 mai4 yu4
67. wo3 ai4 mao1
68. ni3 ai4 yu2
69. ma3 ai4 ri4
70. yu4 wei1 wa1
71. yu4 ao2 wa1
72. yan2 lan3 mao1
73. yan4 ai4 niu1
74. yu4 wei1 yu2
75. yu4 liang2 mao2
76. yu4 wei1 niao3
77. yu4 liang2 ma3
78. mei4 wei1 yan4
79. yu4 ao2 yan4
80. yan4 lan3 miao4
81. ma1
82. ma2
83. ma3

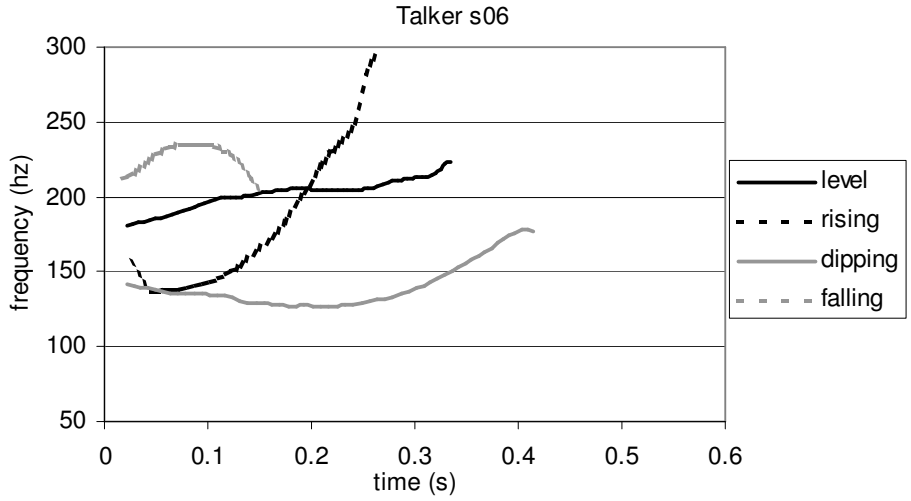
84. ma4
85. lao3 lan2 niao3 liang2 mei3 lv2

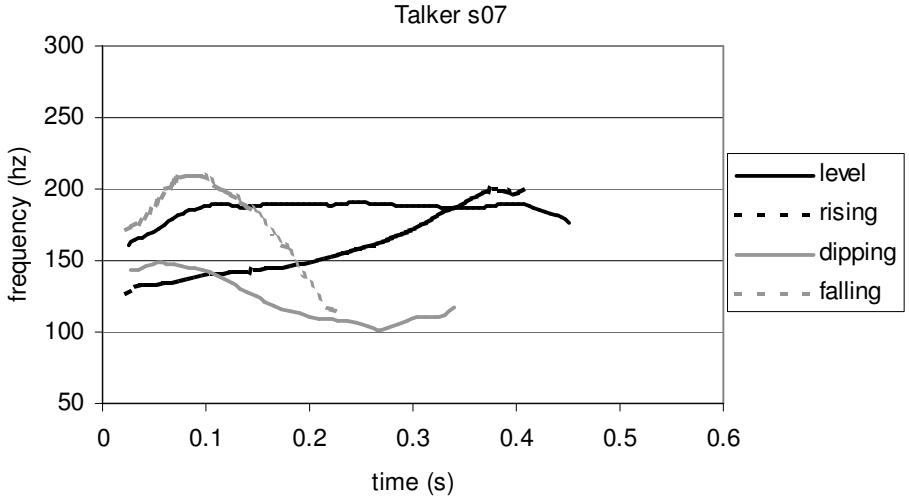
Appendix B: Monosyllabic and Tri-Syllabic Stimuli for Individual Talkers in Experiment 1

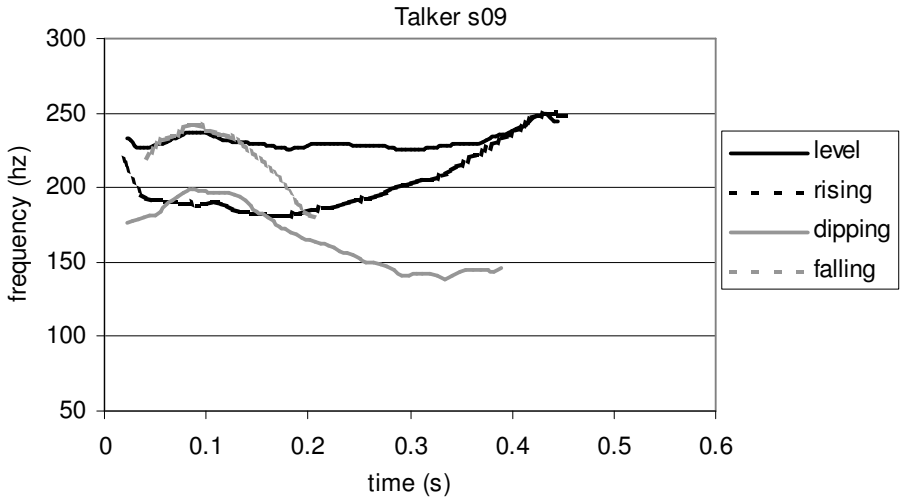


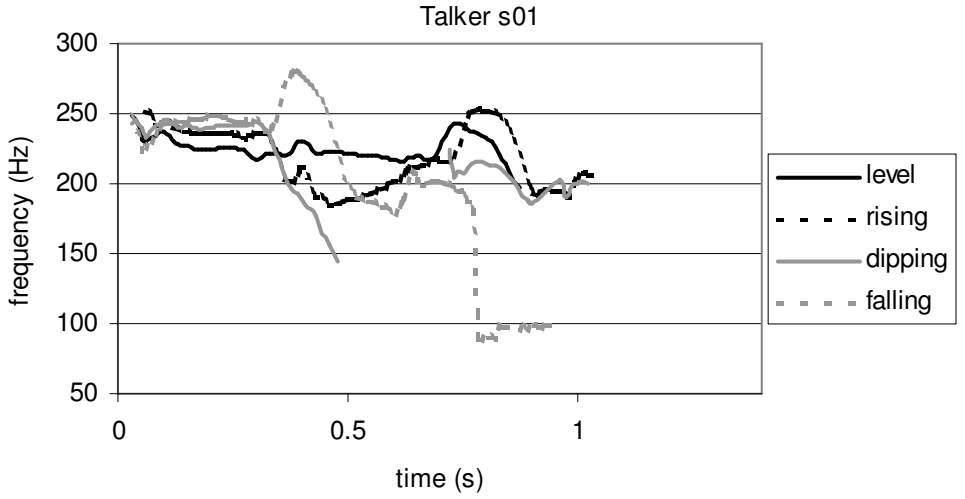


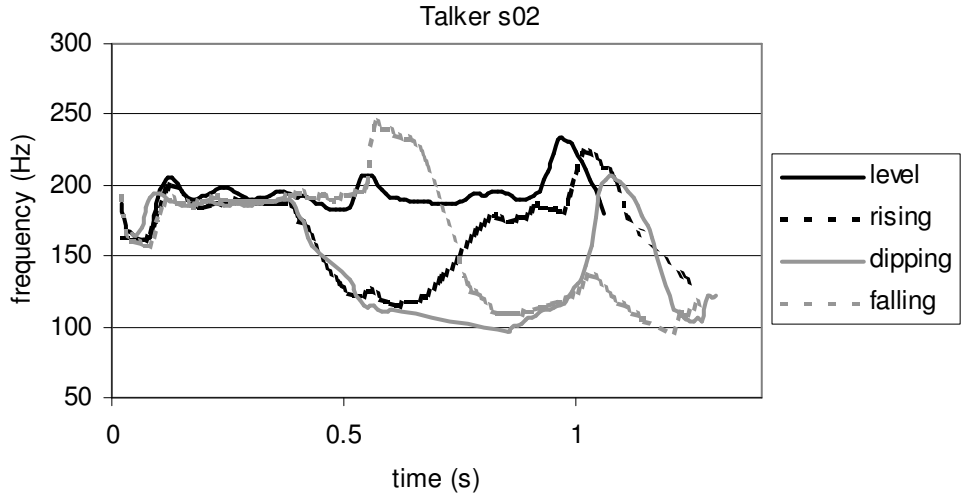


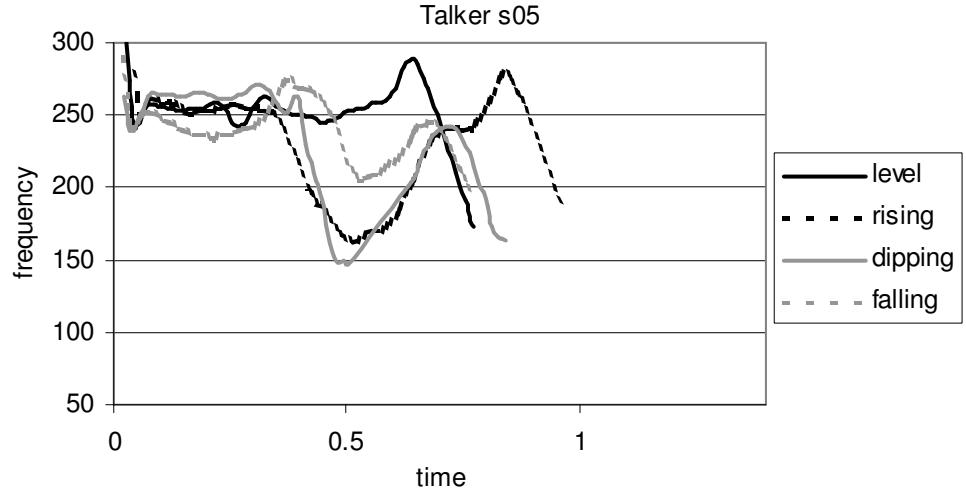

Talker s06
300
250
200
150
100
50
frequency (hz)
0
0.1
0.2
0.3
0.4
0.5
0.6
time (s)
level
rising
dipping
falling

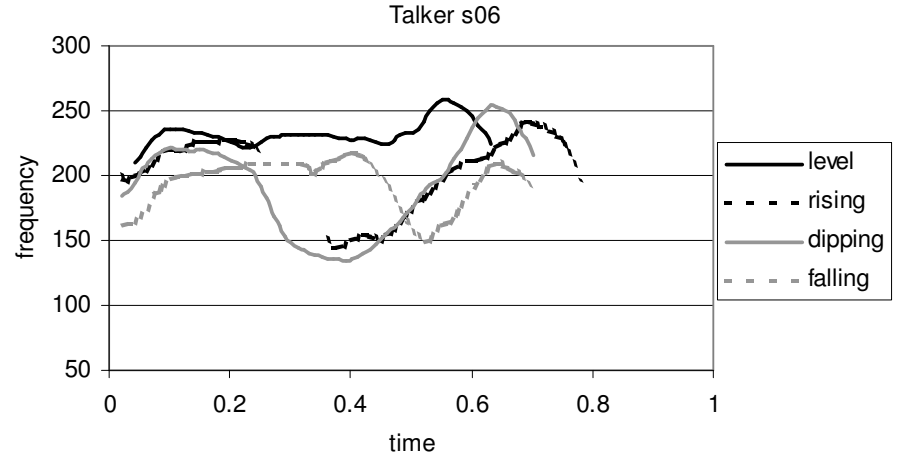

Talker s07
300
250
200
150
100
50
frequency (hz)
0
0.1
0.2
0.3
0.4
0.5
0.6
time (s)
level
rising
dipping
falling

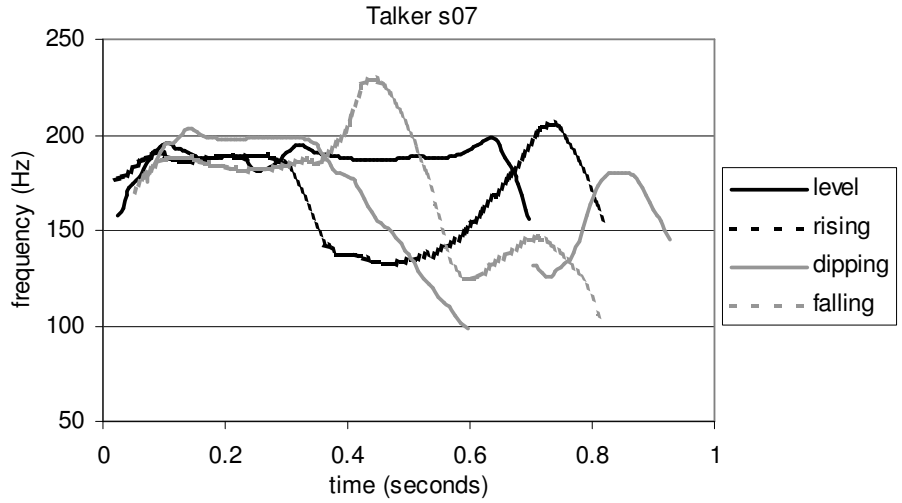

Talker s09
300
250
200
150
100
50
frequency (hz)
0
0.1
0.2
0.3
0.4
0.5
0.6
time (s)
level
rising
dipping
falling

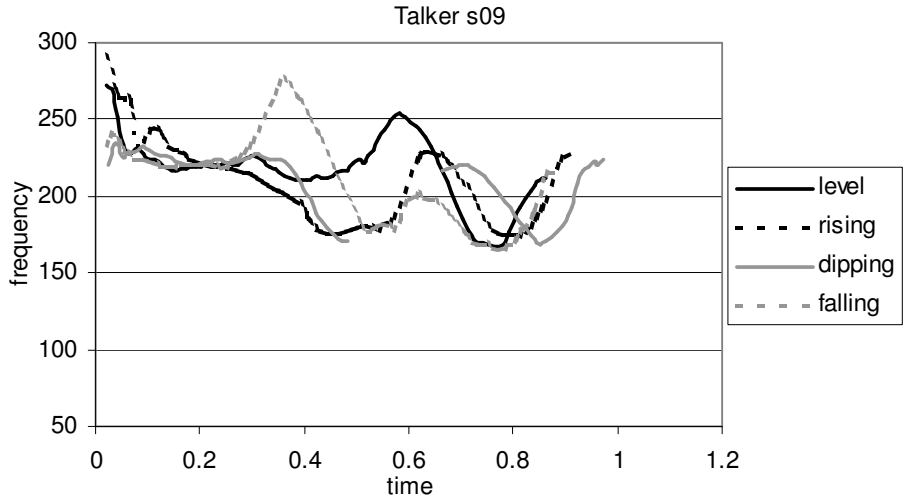

Talker s01
300
250
200
150
100
50
frequency (Hz)
0
0.5
1
time (s)
level
rising
dipping
falling


Talker s02
300
250
200
150
100
50
frequency (Hz)
0
0.5
1
time (s)
level
rising
dipping
falling


Talker s05
300
250
200
150
100
50
frequency
0
0.5
1
time
level
rising
dipping
falling


Talker s06
300
250
200
150
100
50
frequency
0
0.2
0.4
0.6
0.8
1
time
level
rising
dipping
falling


Talker s07
250
200
150
100
50
frequency (Hz)
0
0.2
0.4
0.6
0.8
1
time (seconds)
level
rising
dipping
falling


Talker s09
300
250
200
150
100
50
frequency
0
0.2
0.4
0.6
0.8
1
1.2
time
level
rising
dipping
falling

Appendix C: Instructions for Listeners in Experiment 1

Instructions for discrimination task

For this experiment you will listen to pairs of sounds. These sounds are recordings of people speaking in Mandarin Chinese. Some of these sounds are one syllable and some are three syllables. You will hear 6 different talkers. Your task is to indicate whether you think the members of each pair are the same or different to each other. You should work quickly without sacrificing accuracy. Do not enter your response until the end of the pair of sounds.

This experiment has three sub-parts:

1. Familiarization: In this part, you will hear all the stimuli in the experiment. Your job is to simply listen to the syllables.
2. Practice session: In this part, you will practice the discrimination task. First, you will practice with one syllable pairs and then with three syllable pairs.
3. Experiment: This part is the actual experiment. The experiment has three blocks with a break between each block. In the first half of each block you will listen to the one syllable pairs and then in the second half, you will listen to the three syllable pairs. Your task will always be to indicate whether the members of the pair are the same or different. You will enter your response by pressing the appropriate button on the button box in front of you.

Please ask the experimenter if you have any questions at any time.

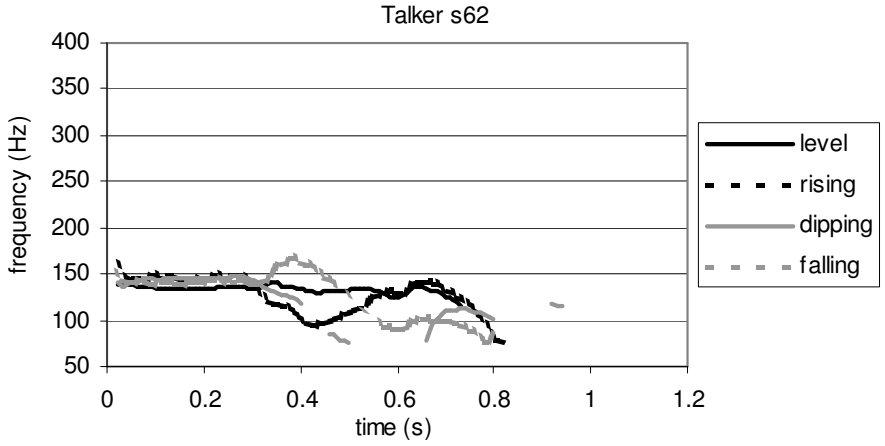
Please let the experimenter know when you are finished with the experiment.

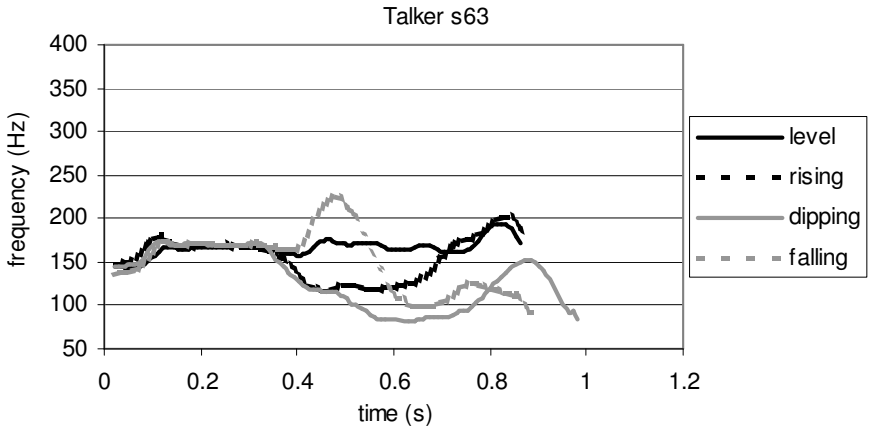
Appendix D: Table of d' Scores for Experiment 1. Standard deviations are shown in paratheses

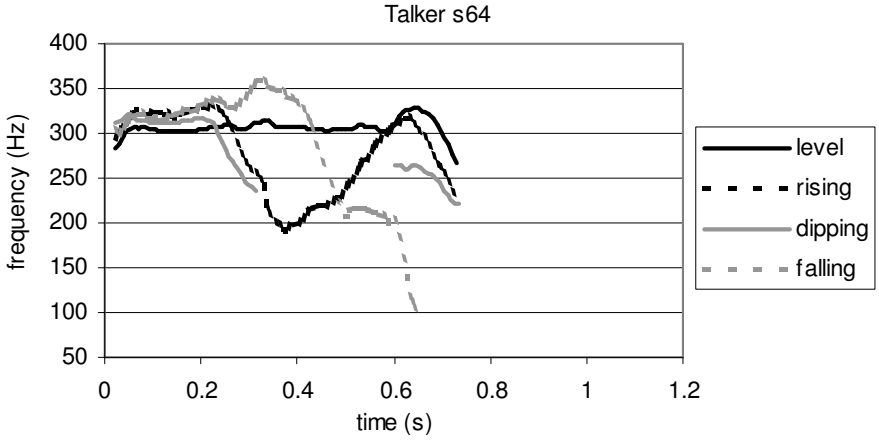
		Tone pair						
		level-rising	level-dipping	level-falling	rising-dipping	rising-falling	dipping-falling	average
Mandarin	monosyllabic	4.13 (0.46)	4.19 (0.46)	4.11 (0.39)	3.96 (0.58)	4.29 (0.40)	4.17 (0.35)	4.14 (0.36)
	tri-syllabic	4.24 (0.69)	4.25 (0.56)	4.23 (0.55)	2.94 (0.34)	4.11 (0.60)	4.27 (0.44)	4.01 (0.48)
English	monosyllabic	3.16 (0.96)	3.96 (0.79)	3.68 (0.92)	3.62 (0.92)	3.54 (1.07)	3.59 (0.84)	3.59 (0.85)
	tri-syllabic	3.56 (0.94)	3.81 (1.01)	3.78 (1.06)	2.47 (0.82)	3.53 (1.05)	3.63 (1.01)	3.46 (0.92)

Appendix E: Stimuli for Individual Talkers in Experiment 2

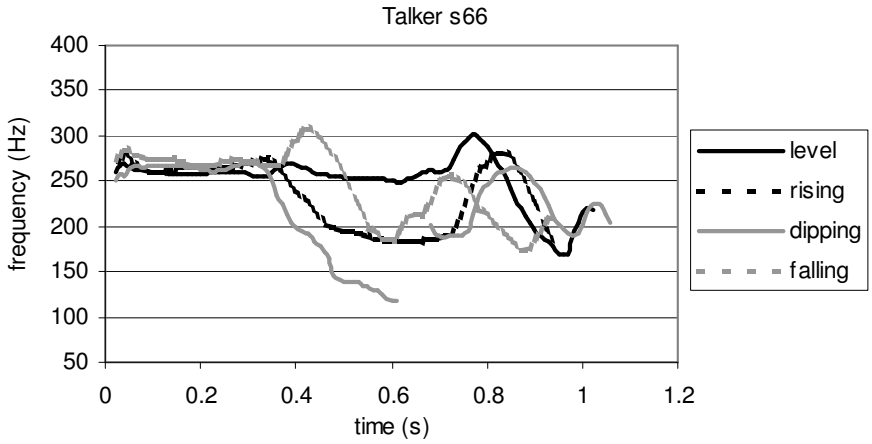
Level – Falling tonal frame



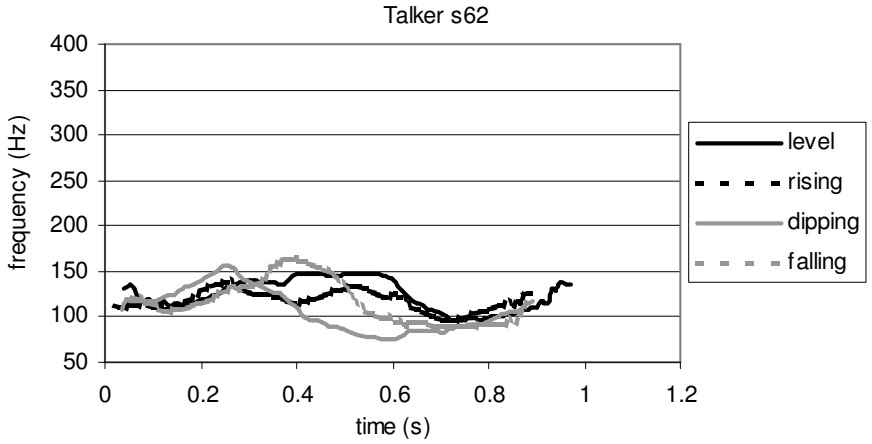


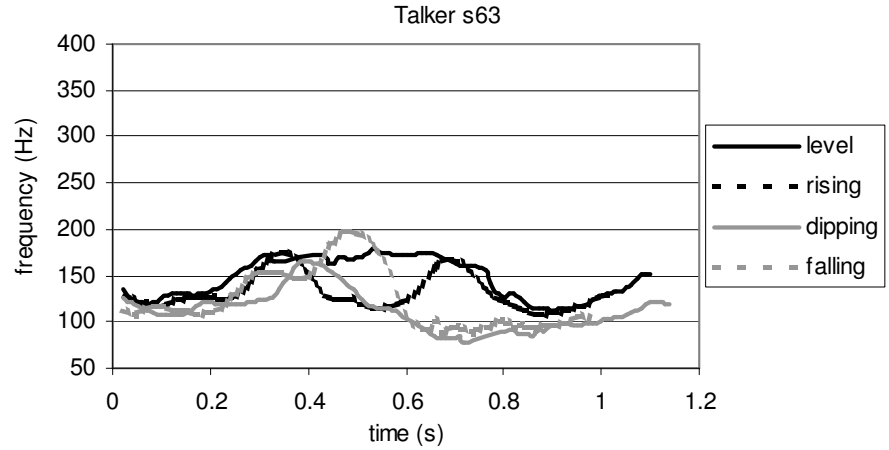


Level – Falling tonal frame continued

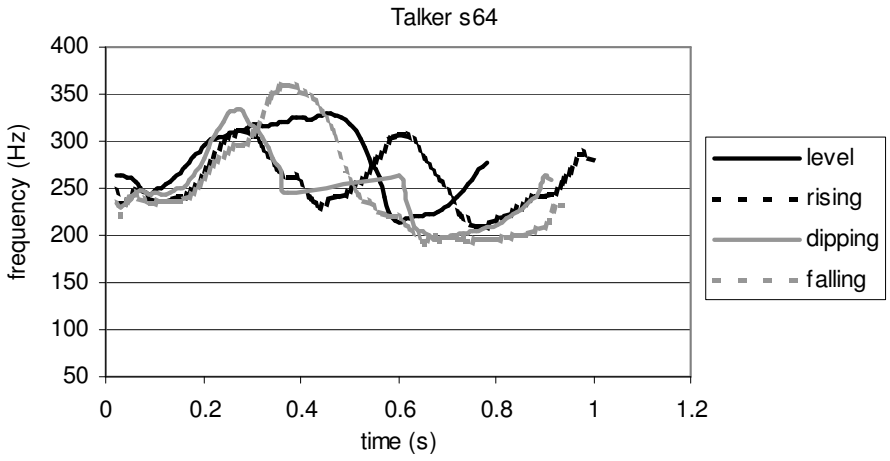


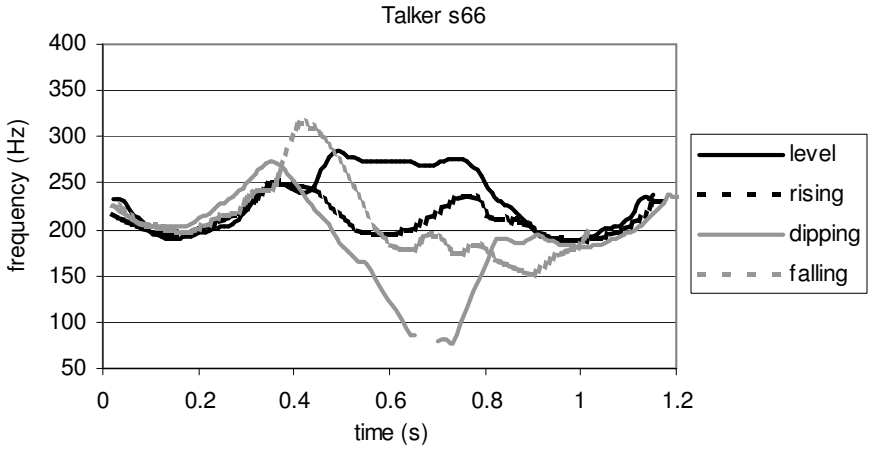
Rising – rising tonal frame



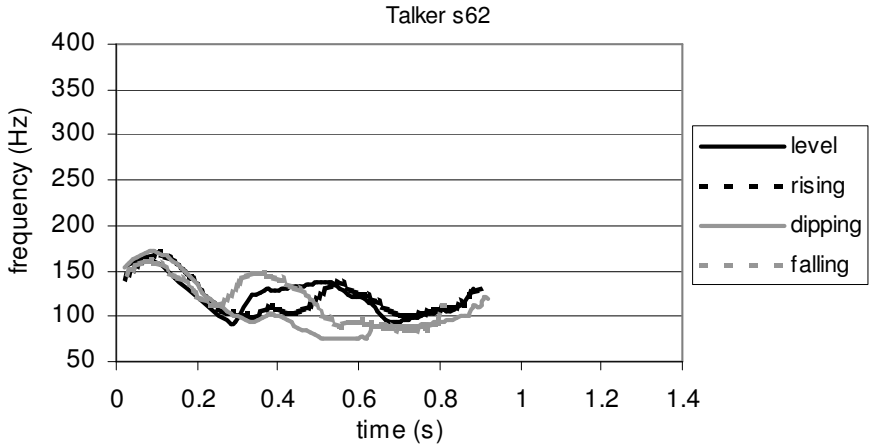


Rising-rising tonal frame continued

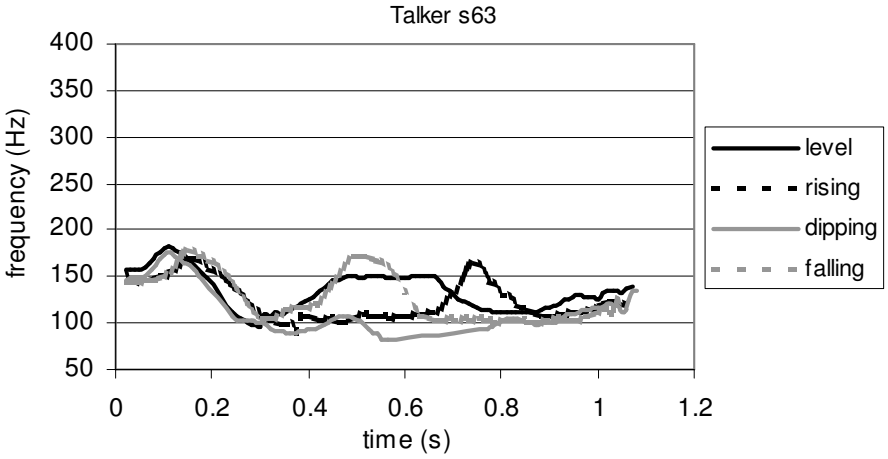


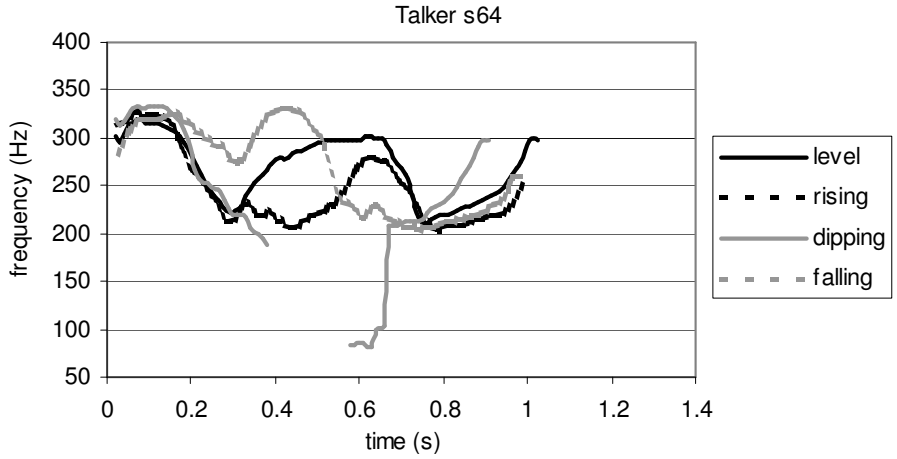


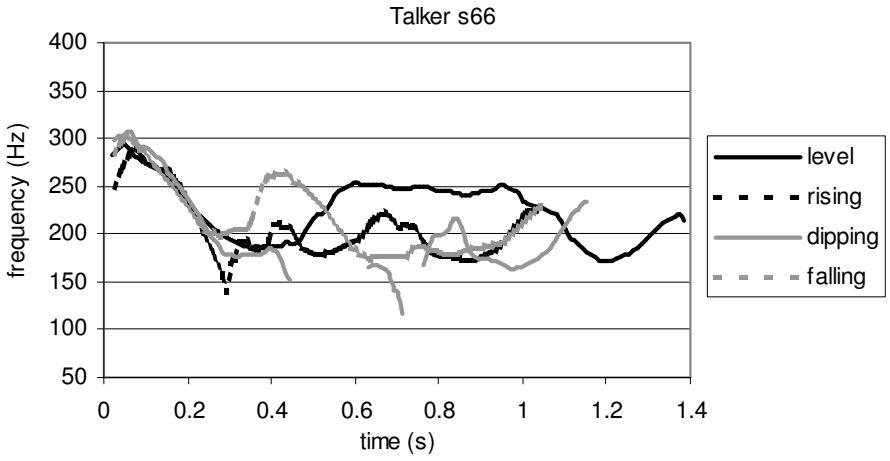
Falling – rising tonal frame



Falling – rising tonal frame continued







Appendix F: Tables of Individual Acoustic Measurements for Experiment 2

Talker s62

Frame	Tone	avg. pitch	max. pitch	% max. pitch	min. pitch	% min. pitch	start pitch	end pitch	% glott.
level–falling	level	133	141	16.6	125	96.4	133	125	0.0
	rising	126	130	98.2	94	37.3	123	130	0.0
	dipping	126	130	17.1	121	17.1	130	glott.	81.3
	falling	127	166	22.3	89	100	140	89	0.0
rising–rising	level	125	148	17.8	118	99.6	136	119	0.0
	rising	118	133	78.1	116	32.2	124	122	0.0
	dipping	110	134	0.3	75	89.1	134	77	0.0
	falling	116	164	21.5	93	98.7	139	93	0.0
falling–rising	level	124	138	0.4	96	0.4	96	123	0.0
	rising	121	137	83.4	98	0.4	98	126	0.0
	dipping	111	101	18.2	76	60.7	94	76	25.3
	falling	119	146	19.9	91	98.9	123	91	0.0

Talker s63

Frame	Tone	avg. pitch	max. pitch	% max. pitch	min. pitch	% min. pitch	start pitch	end pitch	% glott.
level–falling	level	166.4	176	17.1	158	3.1	158	164	0.0
	rising	153.9	179	99.1	116	7.1	118	179	0.0
	dipping	131.7	133	0.6	101	36.9	133	glott.	37.5
	falling	149.6	225	30.9	99	96.2	165	100	0.0
rising–rising	level	145.1	177	32.5	131	98.9	166	130	0.0
	rising	131.1	168	78.4	115	33.3	129	134	0.0
	dipping	113.9	138	0.3	95	37.6	138	glott.	61.6
	falling	123.1	198	24.3	87	84.1	147	90	0.0
falling–rising	level	134.6	151	86.8	108	0.3	108	131	0.0
	rising	124.0	164	89.5	101	20.5	108	144	0.0
	dipping	114.3	107	22.9	92	0.3	92	glott.	66.9
	falling	127.1	171	41.0	102	91.5	109	103	0.0

Talker s64

Frame	Tone	avg. pitch	max. pitch	% max. pitch	min. pitch	% min. pitch	start pitch	end pitch	% glott.
	level	306.0	314	11.1	304	48.1	306	303	0.0
level–falling	rising	276.3	291	99.3	193	29.0	268	291	0.0
	dipping	283.5	255	0.7	236	10.9	255	glott.	87.0
	falling	297.0	359	19.8	206	91.7	329	218	0.0
	level	278.4	330	67.6	253	99.7	308	253	0.0
rising–rising	rising	257.4	307	77.2	230	22.2	262	262	0.0
	dipping	248.3	293	2.2	244	16.4	293	glott.	50.9
	falling	249.0	359	19.8	210	98.7	307	210	0.0
	level	269.9	301	80.9	235	0.5	235	265	0.0
falling–rising	rising	245.6	281	76.8	207	26.1	219	246	0.0
	dipping	240.0	221	0.4	182	33.9	221	glott.	66.1
	falling	267.9	331	31.3	217	90.1	278	228	0.0

Talker s66

Frame	Tone	avg. pitch	max. pitch	% max. pitch	min. pitch	% min. pitch	start pitch	end pitch	% glott.
	level	251.1	270	16.0	250	75.9	255	261	0.0
level–falling	rising	234.1	248	0.6	183	76.1	243	196	0.0
	dipping	221.2	239	0.3	132	51.4	239	189	32.9
	falling	245.0	307	26.0	185	77.4	267	213	0.0
	level	231.5	286	17.8	243	0.3	243	263	0.0
rising–rising	rising	211.1	249	0.2	194	47.4	249	215	0.0
	dipping	200.3	263	0.2	216	14.4	fric.	glott.	62.9
	falling	209.1	314	20.6	175	99.7	242	175	0.0
	level	224.6	253	21.4	204	0.2	204	230	0.0
falling–rising	rising	205.4	221	43.6	178	43.6	186	211	0.0
	dipping	202.2	198	99.1	137	79.2	178	198	52.1
	falling	215.7	263	21.3	175	86.5	205	176	0.0

Appendix G: Instructions for Listeners in Experiment 2

Instructions for discrimination task

For this experiment you will listen to pairs of sounds. These sounds are recordings of people speaking in Mandarin Chinese. You will hear 4 different talkers. Your task is to indicate whether you think the members of each pair are the same or different. The members of each pair will always be recordings of the same person. You should work quickly without sacrificing accuracy. You have a maximum of 3 seconds to enter your response. If you do not enter your response within 3 seconds, the next pair will start automatically. Do not enter your response until the end of the pair of sounds.

This experiment has three sub-parts:

1. Familiarization: In this part, you will hear all the stimuli in the experiment. Your job is to simply listen to the syllables.
2. Practice session: In this part, you will practice the discrimination task.
3. Experiment: This part is the actual experiment. The experiment has three blocks with a break between each block. Your task will always be to indicate whether the members of the pair are the same or different. You will enter your response by pressing the appropriate button on the button box in front of you. Use your right index finger to press the "different" button and your left index finger to press the "same" button.

Please ask the experimenter if you have any questions at any time.

Please let the experimenter know when you are finished with the experiment.

Appendix H: Table of d' Scores for Experiment 2. Standard deviations are shown in paratheses.

		Tone pair						
		level-rising	level-dipping	level-falling	rising-dipping	rising-falling	dipping-falling	Avg.
Mandarin	level-falling tonal frame	3.97 (0.75)	4.04 (0.81)	3.95 (0.65)	2.71 (0.63)	3.90 (0.72)	3.72 (0.59)	3.71
	rising-rising tonal frame	3.94 (0.50)	4.28 (0.51)	4.24 (0.36)	4.22 (0.53)	3.90 (0.59)	4.12 (0.57)	4.12
	falling-rising tonal frame	3.98 (0.71)	4.10 (0.71)	4.03 (0.49)	4.14 (0.69)	3.98 (0.61)	4.09 (0.61)	4.05
English	level-falling tonal frame	3.11 (1.34)	3.05 (1.15)	3.28 (1.01)	2.56 (1.11)	3.27 (1.06)	3.14 (1.15)	3.07
	rising-rising tonal frame	2.78 (0.83)	3.35 (1.01)	3.10 (0.99)	3.15 (0.91)	3.20 (1.06)	3.06 (1.02)	3.11
	falling-rising tonal frame	2.06 (0.80)	3.13 (0.95)	2.88 (1.03)	2.91 (1.10)	2.62 (1.07)	3.02 (0.93)	2.77

Appendix I: Instructions for Talkers in Experiment 3

Instructions for imitation task

In this part of the experiment, you will hear recordings of one talker saying either one syllable or three syllable sequences which differ from one another only in pitch. Your task is to repeat these syllables as closely as possible. Between each syllable/set of syllables, there will be a three second pause. During this pause you should repeat the syllable(s).

Appendix J: Instructions for Mandarin Judges in Experiment 3

Instructions for Mandarin tone identification task

For this task, you will be listening to Mandarin Chinese nonsense syllables. Your task is to indicate which tone (i.e. level, rising, dipping or falling) you think they are trying to say. These productions will either be one syllable, "ra" or three syllables "ra ra ra." For the three syllable stimuli you will be told the tones for the first and third syllables and will identify the tone for the middle syllable.

After you hear each production, you should push one of four buttons labeled in the following way:

rā level	rá rising	rǎ low dipping	rà falling

On each trial, enter your response as quickly as possible without sacrificing accuracy. If, after you enter your response, the next trial does not begin, enter your response again. Do NOT press BEGIN during the experiment.

There will be four blocks of trials within this experiment. After each block, "Take a break now" will appear on your computer screen. You should exit the booth to take a break.

Ask the experimenter if you have any questions at any time. Let the experimenter know when you are finished with the experiment.