The Phonetics of Mandarin Tones in Conversation
⟨汉语会话声调之语音⟩

by

Daniel Scott Brenner

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF LINGUISTICS

In Partial Fulfillment of the Requirements
For the Degree of

Doctor of Philosophy

In the Graduate College

The University of Arizona

2015

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Daniel Brenner, titled “The Phonetics of Mandarin Tones in Conversation” and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: 7 August 2015
Natasha Warner

Date: 7 August 2015
Michael Hammond

Date: 7 August 2015
Miquel Simonet

Date: 7 August 2015
Feng-Hsi Liu

Final approval and acceptance of this dissertation is contingent upon the candidate’s submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 7 August 2015
Dissertation Director: Natasha Warner

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Daniel Brenner

Acknowledgements

It is a daunting task acknowledging the countless individuals that have together made this dissertation possible... It truly does take a village to raise a doctoral student.

First, I would like to acknowledge the many that will not find their names here. Inevitably, the random associative process that is taking place in my brain will leave some holes in its path, where branching decisions took it away from some semantic area and never led back to it. For this I can only offer my apologies for the omission. I appreciate all the unnoticed giving that has occurred.

Those mentions that do follow are in no particular order of importance, but only of my train of thought.

Natasha Warner, Mike Hammond, Miquel Simonet, and Feng-Hsi Liu have been a spectacularly supportive and constructive committee. I apparently chose very well, and it has made all the difference. This dissertation owes much of its coherence and scholarly utility to their guidance. They have given a broken, dissatisfied, self-doubting student the experience of success, despite my limitations. I might, somehow, actually be able to do this stuff, huh?

LI Qinzhuo was an extremely reliable research assistant on the project for a summer, helping to recruit participants, and ran the experimental set-up for many of them while I was away in Beijing. She also contributed gold standard transcriptions for the dictation task, and performed a certain amount of token identification for use in stimuli. I am very happy to have had the benefit of her generous help.

CHEN Yan helped with task piloting, transcriptions, instruction translation, and plenty of lively discussion about tones.

Andrew Lotto and Nico Carbonell have become indispensable colleagues, great friends, and much needed research consultants. I greatly value the outside perspective that they have brought to me, in so many ways, in research and just sitting on the patio.

My mother, Ann Brenner, like most mothers perhaps, deserves so much more gratitude than I have ever given. She has provided absolute support in nearly everything I get into, and introduced me to so many things that are a staple in my existence. I delight in having good people around me, and she has always been a steady, venerable fixture of kindness for me.

Susan Brenner and Lyle Decker, my Arizonan family, have always made me feel at home with them, taking me hiking, cooking good food, and commiserating through difficult times. They also let me store my jalopy motorcycle in their garage without excessive ridicule or sarcasm. Thank you for putting up with me and giving me such a relaxing environment to get some distance from my sources of stress.

David Brenner kept me focused on the more mad-scientist opportunities of my graduate studies, and though we are poor at keeping in touch, when we connect it is always a happy thing for me. I hope that we can do it more often now that this chapter is behind me.

Ken, Yvonne, and Mariko Brenner, and Vita Kawashima, have shared so many wonderful holidays with me, and some unfortunate ones as well, but are generous and kind beyond words, and always make me feel like a hopeful child, teaching me to bake pies, assemble a large-format camera, and endlessly piling the most fascinating books around me. I have such fond memories of walks in the snow, mountains of home-grown heirloom tomatoes, and whimsical stories told around the kitchen table.

Jessamyn Schertz was an ever supportive, and terribly “with-it” office mate, who traded jobs with me twice, and always tried to encourage me when I was struggling. She has also been a wonderful academic model to have around.

Alan Hogue, Bryan Gordon, and Jaime Parchment were my social outlet for much of graduate school. I have deeply enjoyed our playful happy hours and evenings, and hope that we will find a way to continue gathering.

Lee Sechrest and A.J. Figueredo welcomed me into their fold of data and methods researchers, and I have benefitted greatly from the Psychology group they organize. The discussions there are unlike any I have experienced in other departments, and opened my eyes to many new and exciting themes and questions.

B.J. Berardo toughened me up as a teenager, and came to the rescue with the Polish nasal vowel example found in the introductory chapter. Good lookin' out, brother!

This dissertation also benefitted from the generous support of NSF EAPSI grant OISE-1311085, a summer stipend and lodging in Beijing from the Chinese Science & Technology Exchange Center, and RA support and subject compensation funding from University of Arizona SBSRI grant 14DRS0268. Considerable support was also given by Dr. HU Fang at the Chinese Academy of Social Sciences in Beijing during a summer research trip.

My good friends Mike Bindschadler, Bryan Naegele, Eric Brown, Moon-Ho Hwang, and Mars Kim were vital sources of support during graduate school. I could not have survived without you.

Jae-Hyun, Stan, and Shelley, thank you for creating a home with me, and for sharing the ups and downs of life. I am grateful to have such peaceful and loving companions.

This dissertation is dedicated to my father, Ronald Brenner, who has fought for me so ferociously and supported me so steadfastly all these years, even through times I must have made him wonder why he ever bothered, and all without even the most casual of thanks or acknowledgments. There is no question I could not have made it to/through grad school without his home, his books, his love of learning, his skeptical unorthodoxy, and just his care and attention. I could not have hiked this trail and seen these views of the world without the bridges he built for me and somehow kept me from burning.

Table of Contents

Table of Contents—*Continued*

Table of Contents—*Continued*

List of Tables

List of Tables—*Continued*

List of Figures

List of Figures—*Continued*

List of Figures—*Continued*

Abstract

Mandarin tone categories are universally thought to center on pitch information, but previous work (Berry, 2009; Brenner, 2013) has shown that pitch cues reduce in the conversational context, as do the other concurrent cues such as duration or intensity that secondarily signal tone categories. This dissertation presents two experiments (an isolated word perception experiment, and a dictation experiment) aimed at discovering how Mandarin listeners deal with these reduced cues under everyday conversational conditions. It is found that detailed spectral information is far more useful in the perception of Mandarin tones—both in isolated words and in the perception of full conversational utterances—than pitch contours, and that the removal of pitch from the recordings does not greatly influence perception of the tone categories.

Chapter 1

Introduction

1.1 The Complexity of Conversational Speech Sounds

The study detailed in this dissertation addresses how Mandarin listeners perceive tones in conversational speech. Tones, and the perception of conversation in general, relate to the human psychological system which manages the innate capacity for speech.

A very common conception of the functioning of speech as a communication device is that a sequence of units referred to as speech "sounds" such as consonants and vowels are employed as a kind of code by which speakers and listeners are able to convey meaningful larger units such as morphemes, words, utterances, turns at talk, and so forth. For example, the words 'coat' and 'wrote', 'sip' and 'ship', 'pout' and 'peat' are different words in English, and the sequence of sounds that make them up, along with prosodic characteristics such as stress and length, indicate which is intended. These speech sounds are themselves each composed of a set of "features", articulatory or acoustic properties which allow language users to distinguish them from one another. For example, ‘nip’ and ‘dip’ are distinguishable because in ‘nip’, the first sound [n] is *nasal*, produced with the velum lowered to open the port into the nasal cavity producing a split-chambered acoustic resonating space, while the first sound in ‘dip’ [d] is *oral*, meaning the nasal port is closed, and the oral passage through the mouth is the only resonating space above the larynx. Otherwise, the two initial sounds are produced nearly identically, but this difference alone is sufficient for listeners to distinguish the two sounds, and therefore, the two words.

These *features* which distinguish speech sounds may themselves be sequential in time (such as a silent closure followed by an aspirated release burst in voiceless plosives like the *p* [p^{h}]) in ‘pin’ or simultaneous (such as the first two formants in vowels like the *ee* [i] in ‘feed’ or the *o* [o] in ‘poke’), but each sound has a specified set of features which designate it and separate it from the other sounds. Speakers aim to produce the features to encode a message, and listeners listen for them in order to decode the sequence of sounds and recover the message. This dissertation, in broad terms, expands on the existing work on how speech sounds and the features used to designate them are affected by speech style—in this case, the difference between their acoustic and perceptual properties in careful speech, like when reading a list of words aloud, and casual conversation, chatting with a close friend or family member.

A great deal of the history of phonetic research has aimed to catalog the cues needed to convey any given speech sound (notable early works are Steinberg, 1934; Potter, 1945; Kopp and Green, 1946; Cooper et al., 1951, 1952; Ladefoged and Broadbent, 1957; Halle and Stevens, 1962), e.g. how a speaker conveys to a listener that 'sprite' is intended rather than 'sprout', and detailing the contexts which affect how those cues are transmitted and received. For example, in very noisy environments or under other difficult listening situations like during a noisy concert, one may produce the sounds in an exaggerated manner, for example "Do you want to go?" [duː juː wɔnt tu goː], while under normal circumstances to a close friend, one might produce something closer to "Yauna go?" [jɔːnə goː].

Conversational speech research, however, suggests that listeners cannot simply be recognizing one sound after another and piecing them together into words like "beads on a string" (Bloomfield, 1933; Öhman, 1966; Fant, 1973; Kent and Minifie, 1977; Fowler, 1984), listening for 'I', 'm', 'g', 'o', 'i', 'ng', 't', 'o'…because they are quite likely to hear a large variety of acoustic versions—"I'm gonna", "Imana", "Ima", “Ama”, etc. The picture that is emerging is that speech sounds may be transmitted to listeners with a tremendous variety of acoustic versions (Greenberg, 1999; Ernestus et al., 2002; Pluymaekers et al., 2005a,b; Ernestus and Warner, 2011), including some which simply delete some sounds or syllables from the signal altogether (Johnson, 2004; Schuppler et al., 2012). For example, the popular colloquial contraction “aowanna” [aʊwaːnə] (‘I don’t want to’) contains only the remnant “ao” [aʊ] to signal the entirety of “I don’t” aɪ dont. And sounds are not only produced in common, almost standardized shortened forms like this one has become, but may appear in unexpected, unpredictable variations, such as ‘yesterday’ produced as “yeshay” [jɛʃːeː] (Ernestus and Warner, 2011, p.) or ‘what do you’ sounding like ‘wuhya’ [wʌːjə].

Despite the wild variation in the way sounds are produced and combined, conversation produces little comprehension difficulty under normal circumstances. This is partly because of information available from the context of the utterance (“We’re in the grocery store produce section, so he probably said ‘three bananas’ rather than ‘free the Nannas!’ ”), and partly because the production of a speech sound cue has side-effects which themselves also become cues to listeners who know where to “look”. Listeners do indeed have great facility in monitoring many channels concurrently

(Holt and Lotto, 2006; Chandrasekaran et al., 2010; Schertz, 2014) and, at the same time, are aware of how likely words are in a given context. They utilize information about what words can go next to one another syntactically, e.g. “in the…” may be followed by “…morning / West / bottom / mind…” etc., but cannot be followed by “…of / because / buy / shorten”, and so on. Also, in certain real world environments some words are more prone to appear than others, for example “Hand in your…” will most likely be followed by “…homework / exam” etc. at school, but by “…report / timesheet / spreadsheet” etc. in a business setting, or by “…badge and gun” in the case where a police officer is being fired, and listeners capitalize on these correspondences also. At the same time, they monitor how fast the speaker talks, how long consonants and vowels are compared to others near them, and make use of information about segments present in their neighboring segments. For example, a listener may hear nasalization on a vowel signalling a following nasal consonant even if the consonant itself is not clear.

Humans are confined in speech communication to the kinds of sounds they can produce with their vocal tracts and the properties of sounds that anyone can make and anyone can hear. But there remain many different “channels” in the acoustic signal that speakers use to convey information, some which are tied together to a large extent like the formant deflection on the ends of vowels caused by stop consonant constrictions (Story and Bunton, 2010), and some which can be produced more or less independently and concurrently with others. Voicing or intonation are this latter kind, where one can do “Oscar’s here!”, or “Oscar’s here?” with the same segments but different intonational patterns overlaid on the words. Then there are channels in between like nasalization which interacts with some but not other channels of acoustic information. E.g. vowels can be nasalized as in the Polish ‘sąd’ [sãːt] “court” vs. ‘sad’ [saːt] “orchard”; and stops can be nasalized, e.g. ‘Mom’ (both stops are nasal) vs. ‘Bob’ (both stops are oral); but oral fricatives are not reliably nasalized even when speakers of the language think of them as nasal, as in Scottish Gaelic (Warner et al., 2015).

One such channel that can be used more or less independently, as mentioned above in the intonation example, is pitch. All languages make ample use of pitch to convey various kinds of information including attitudinal distinctions like favor or skepticism (genuine vs. sarcastic “That’s great!”), or to differentiate speech acts like statements and questions (“It’s made of seaweed” vs. “It’s made of seaweed?”), but an estimated 41.8% also make use of pitch to differentiate words or syllables themselves (Maddieson, 2011). In a “tone language” such as this, two or more words may be pronounced with identical consonants and vowel sounds, e.g. the Mandarin Chinese syllable [wɛn], but be distinguished by the pitch of the voice used when pronouncing the words (Fromkin, 1978; Yip, 2002). These pitch patterns are called “tones”, and they are part of the specification of a word in tone languages, such that when a speaker learns a new word, they must learn the correct pitch pattern for the word along with the correct syllables. In Mandarin, pronouncing [wɛn] with a steadily falling pitch through the syllable means ‘to ask’ and is written ⟨问⟩, while the same syllable ([wɛn]) with a dipping-then-rising pitch means ‘to kiss’, and is written ⟨吻⟩. The contribution of tones (pitch patterns) to the system of sounds in Mandarin words has been quantified, and the number of words distinguished by tone in the language (e.g. ‘to ask’ and ‘to kiss’ as above would become identical-sounding words, [wɛn]) was found to be as great as the number of words distinguished by vowels (e.g. [lan] with rising tone, ‘blue’, and [lin] with rising tone, ‘forest’, would sound identical [l-n] with rising tone, if there were no vowel categories) (Surendran and Levow, 2003).

As mentioned, the way segmental sounds are produced in conversational speech have been found to differ greatly from the “beads-on-a-string” conception of speech, and recently there has been a great deal of interest in conversational or “spontaneous” speech (Miller et al., 1984; Ernestus et al., 2002; Gahl, 2008; Kuperman et al., 2008; Warner, 2012), its variability (Johnson, 2004; Tucker and Warner, 2007; Ernestus and Warner, 2011; Schuppler et al., 2012), and how speakers and listeners manage the highly multidimensional realizations of speech sounds (Holt and Lotto, 2006; Schertz, 2014). The pitch patterns which contribute to the definition of words in tone languages are in some ways similar to segments. Liu and Samuel (2004) conclude that they are realized flexibly and dynamically, so that for example when whispered, speakers will use other cues to realize the tones. In some ways, however, tones are quite different from segments. Cutler and Chen (1997) show that tone perception is more error prone, and occurs more slowly than segmental perception. It is of some interest, then, how tones will be realized in conversation, whether the same kinds of surprises as in segments present themselves in tones, and how listeners cope with conversational variability.

Although considerable bodies of work have explored the production and perception of tones in Mandarin and other languages in carefully spoken word lists (Kiriloff, 1969; Howie, 1976; Cao, 1992; Cutler and Chen, 1997; Xu, 1997; Kong and Zeng, 2006; Cabrera et al., 2014, and many others), little research has so far addressed how these tonal categories are transmitted or decoded in everyday conversation. For consonants and vowels, it has been found that in many cases, the acoustic information we use to perceive these sounds is less robust in casual conversation than in careful speech (Ernestus and Warner, 2011; Warner, 2011), so that cues that characterize a sound in careful speech may be reduced or absent in conversation. E.g., during conversation, an English speaker may neglect to completely contact the roof of their mouth during the [k] in the word 'weekend', resulting in a word that has noise or only a slight weakening of acoustic intensity during the [k] relative to its surrounding vowels, leading to a word that sounds more like 'weehend' or 'we end'. These same cue reduction processes (whereby cues are less prominently produced than in careful speech) have been reported in many other languages, including Mandarin (Ernestus and Warner, 2011; Tseng, 2004b; Cheng and Xu, 2009, 2013, 2014). Extreme cases where multiple syllables are contracted into single syllables in Taiwanese Mandarin were the topic of a recent dissertation (Cheng, 2012). The project detailed in this present dissertation will extend the scholarship on the reduction of consonants, vowels, and syllables, to the tone categories utilized in Mandarin, examining what happens to tones in the conversational context. These are especially important in language teaching (where conversational reduction often poses an extreme challenge to learners), in understanding how native tone language speakers produce and learn other languages, and in developing more natural and intelligible human language processing systems such as automatic speech recognition and synthesis systems.

1.2 Mandarin Tones

Mandarin is an excellent language for the study of lexical tone. In addition to native speakers being readily found, the tone system has a rich set of various properties that make tone systems interesting (Fromkin, 1978; Yip, 2002; Maddieson, 2011), though there are some properties of African tone systems and Asian tone systems that appear to be mutually exclusive (Chen, 2012). For example, African languages almost exclusively utilize level tones, while Asian languages like Mandarin make heavy use of contour tones; while African language show stepwise processes that shift its level pitches with respect to one another, Mandarin tone combinations are almost entirely independent, such that any of the tones can occur in nearly any syllable. It also exhibits categorical (or near-categorical (Cheng et al., 2013)) swapping of tones in a small number of environments. Mandarin has level and contour tones (where the pitch during the syllable changes in some definitional way), full and "neutral" tones, stress, tone Sandhi, voice quality cues, and so forth. And all of this occurs in a system of 4 full tones, and the neutral tone (sometimes described as "toneless", a tone which occurs in unstressed syllables and whose pitch pattern is largely determined by its neighboring tones (Cao, 1992; Chen and Xu, 2006; Lin, 2006)), a distinction which itself merits interest. The tones have come to be identified by native speakers as simply "tones 1-4" (or sometimes the "high tone", "rising tone", "dipping tone", and "falling tone" respectively), and "the neutral tone" or "clear tone", but for purposes of this dissertation, either the IPA tone symbols (˥, ˧˥, ˨˩˦, ˥˩) and the custom neutral tone symbol "Φ", or the more descriptive names 'high level tone', 'rising tone', 'dipping tone' (a low falling, then rising pitch pattern), 'falling tone', and 'neutral tone' will be used. In Pinyin romanization, the tones are designated by diacritics (¯, ´, ˇ, `, and no diacritic for the neutral tone) above the syllable nucleus letter, where the diacritics are iconic of the pitch pattern to be produced during a syllable of that tone, much like the IPA tone symbols in simplified form. Table 1.1 demonstrates the usefulness of tones in Mandarin by providing a set of words which would be homophonous without tone distinctions.

Characters	Gloss	Pinyin	IPA	Tone #	Tone Name
哥	Elder brother	gē	[gɤ˥]	1	high level
格	Pattern; frame	gé	[gɤ˧˥]	2	rising
哿	Excellent; happy	gě	[gɤ˨˩˦]	3	dipping
虼	Flea	gè	[gɤ˥˩]	4	falling
个	(quantifier)	ge	[gɤΦ]	0	neutral

Table 1.1: A set of words distinguished by tone. The neutral tone phonetic symbol is not an official symbol of the IPA; it was created for use in this dissertation for clarity.

Tones are widely believed to be predominantly distinguished from each other by the pitch contour produced over the voiced sonorants of a syllable (Chao, 1965; Duanmu, 2007; Lin, 2007). Schematic pitch contours for each of the four full tones are shown in Figure 1.1. This is a typical teaching material for the tones, for native children and also second language learners, and shows roughly what the pitch of the voice is meant to do in producing the tone, where horizontal dotted lines indicate a 5-level division of the speaker's normal speaking pitch range. For example, tone 1 is a high level tone, where the speaker maintains a relatively high pitch throughout the tone 1 syllable. During tone 3, the speaker begins mid-low, dips down and then back up to mid-high. In terms of distinguishability, the high level tone and falling tone are readily distinguished from the other tones and each other (even when the pitch space is artificially compressed to a range of just 2Hz (Klatt, 1973)), while the rising and dipping tones are frequently confused (Blicher et al., 1990; Shen and Lin, 1991).

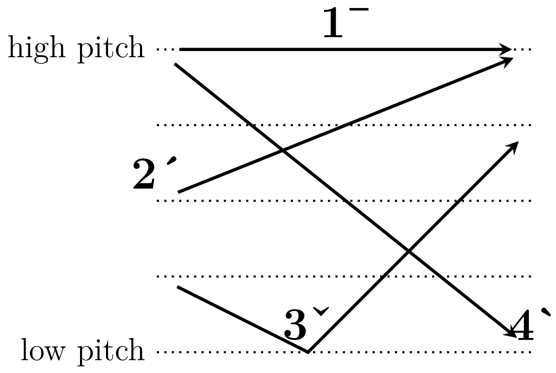


Figure 1.1: Schematic of the standard pitch patterns of tones 1–4 as conceptualized by native speakers, and taught to children and L2 learners. The neutral tone (not shown) meanwhile, is relatively short and varies widely in its pitch realization according to its neighboring tones (Cao, 1992; Chen and Xu, 2006). Iconic Pinyin diacritics are shown next to the tone numbers. Horizontal dotted lines represent a 5-level division of a speaker's normal speaking pitch range. This visual scheme was the basis of the current IPA standard notation for tones, devised by Chao (1930). This figure is repeated in the chapter introductions for reference.

Mandarin exhibits tone Sandhi (Shih, 1997; Chen, 2000; Cheng et al., 2013), a class of phonological process governing the pitch patterns of tone, which in Mandarin neutralize certain tones in some tonal contexts, unlike the tone processes in African languages, which overall have more constrained tone patterning (Yip, 2002). As the most prevalent example, a tone 3 followed by another tone 3 is realized as a tone 2 (or very similar to it (Cheng et al., 2013); it is not known whether listeners are able to make use of the marginal acoustic differences detected in the Cheng et al. paper). The process is cyclical, such that a string of tone 3 syllables will result in a string of tone 2 syllables with a single tone 3 at the end, though this appears to be confined to the prosodic phrase (Shih, 1997).

In addition to the pitch patterns during Mandarin syllables, the duration of syllables (Blicher et al., 1990; Chang and Yao, 2007; Cabrera et al., 2014), how the intensity changes over the syllable (Whalen and Xu, 1992; Liu and Samuel, 2004), what the voice quality is (Yu, 2010; Bissiri et al., 2014), and some properties of the segments themselves (Xu, 1997; Hu, 2004) also provide information about the tone category of a syllable, just as segmental categories can be signalled by multiple correlated cues, e.g. duration, F_1 onset lag, and pitch at onset of voicing can all signal whether a stop is voiced (as in the [d] in ‘dour’) or voiceless (as in the [t] in ‘tower’) (Schertz, 2014), yet tones are more slowly processed and tone category identification in listeners is more error prone than segment identification (Cutler and Chen, 1997). The experiments of this project will examine how tones are realized and perceived in conversation in terms of the cues that are expected from the careful speech research, and compare the combination of available information channels to what is known for segments. This will expand our understanding of conversational speech beyond segments into other channels of information transmission, and also expand our understanding of Mandarin, tone, and language itself.

This dissertation reports on two experiments. One is a tone category identification experiment described in Chapter 2. The other is a dictation experiment investigating comprehension of Mandarin conversational speech, presented in Chapter 3.

1.3 Experiment 1: The Perception of Mandarin Tones in Isolated Words

The first experiment measures how well Mandarin speakers are able to identify tones in read and conversational words taken from separate recordings of a Beijing-area native Mandarin speaker, and measured how much poorer they are at identifying tones when pitch information, or information about which consonants and vowels are present, was artificially removed from the recordings.

In order to separate the contributions of pitch and segmental detail in the perception of tone in isolated words, stimuli for this experiment were resynthesized from the production recordings to manipulate what information listeners have available to identify the tone in the first syllable. Stimuli were careful or conversational speech, and were presented in three forms, either resynthesized using Linear Predictive Coding (O'Shaughnessy, 1988), or "hummed" using an artificial voicing model:

Full: Fully resynthesized[1] so that all information is available; e.g. [bi (dipping) . jiao (falling)] 'compare' sounds like [bi (dipping) . jiao (falling)].

Hum: The first syllable is replaced by a humming sound which has the same pitch as the original syllable; e.g. [bi (dipping) . jiao (falling)] sounds like [*hum* (dipping) . jiao (falling)].

Whisper: Pitch information is stripped from the first syllable to create a "whispered" effect[2];

e.g. [bi (dipping) . jiao (falling)] sounds like [bi (*whispered*) . jiao (falling)].

Fifty Mandarin native listeners heard words in the three conditions described above, and for each word pressed a button to indicate which tone (tone 1, 2, 3, or 4) they heard in the first syllable of the word (e.g. they hear [bi(dipping).jiao(falling)],

[1] All stimuli were resynthesized so that they are comparable in terms of this manipulation.

[2] This is not the same as natural whisper, where it is thought that speakers may exaggerate secondary or other acoustic cues to facilitate tone recognition, though findings are so far inconclusive (Chang and Yao, 2007; Li and Guo, 2012).

and press the “tone 3” button because the dipping tone in the first syllable is known as “tone 3”).

This determines how much information listeners can get from each type of sound, and establishes how much harder it is for listeners to recognize words with each type of information missing, e.g. with no information about consonants and vowels in the first syllable (like [*hum* (dipping) . jiao (falling)]) or without pitch information (like [bi (whispered) . jiao (falling)]), relative to words with all information present (like [bi (dipping) . jiao (falling)]).

1.4 Experiment 2: Transcribing Conversation With and Without Pitch

This experiment measured the contribution of tone to comprehension in whole utterances of Mandarin conversational speech. Even though Mandarin speakers *can* identify tones, that does not mean that they routinely *do* use tones to help them identify which words they are hearing in real conversational contexts. For example in English, listeners may be able to hear the difference between similar words like “prints” and “prince”, but it is unlikely that they need to hear the difference in real contexts, since there are few contexts that both words could occur in. Experiment 2 determined to what extent Mandarin speakers actually need pitch information while listening to Mandarin conversational speech, whereas Experiment 1 investigated how well they *can* perceive tone with or without pitch information available.

Fifty native Mandarin speakers performed a transcription task[3] on the sentences prepared to provide pitch information (fully resynthesized utterances with pitch information intact) or withhold it (artificially "whispered" utterances in which pitch information has been removed). This determines whether listeners are actually using the pitch information. If the conversation context is adequately constraining (e.g. only a small number of nouns can occur after "Please turn in your…" in a given context), tones may not actually be crucial in conversational Mandarin. If this is the case, second language learners and instructors of Mandarin may wish to concentrate their study time in other areas (as tones are notoriously difficult to acquire for speakers of non-tonal languages), and speech technologists may wish to allocate ASR and synthesis development resources differently, depending on their goals.

[3] The tones are explicitly taught in school, and Mandarin listeners' ability to do this task is related to the high degree of literacy and metalinguistic knowledge that results from the educational system there.

1.5 The Big Picture

As a whole, these experiments will provide a picture of the role of pitch in Mandarin perception, both at the level of isolated words, and at the level of full utterances in conversation. The perception experiment reveals how Mandarin listeners make decisions about tone category when provided with different information sources, and how informative each of these sources is. The dictation experiment shows how much pitch information contributes to the comprehension of Mandarin utterances in a more natural conversational context. The combination will also tell us something about how perception at these two scales is related to each other, particularly to do with the contribution of pitch, revealing how much of utterance perception can be explained by the perception of individual words, and how much appears to exist only at the larger intervals. The relative increase in error rates in isolated words or in conversational utterances suggests that pitch plays a greater role at that scale. As a first step, however, it is important to take stock of the role of pitch and segmental information in the recognition of words themselves.

Chapter 2

Partitioning Information for Tone Perception

2.1 Introduction

Speech sounds are thought of as units to be encoded by speakers and then decoded by listeners. A considerable body of work has been conducted over the last several decades into the acoustic cues that speakers make use of in signalling the various speech sounds, and in recent decades quite a bit has been explored regarding how those cues are employed in concert to decode the intended sounds under certain kinds of speech conditions (noise, unfamiliar speakers, unfamiliar dialects, etc.). With the need for experimental control of variables such as speech rate, segmental and prosodic contexts, and so on, nearly all of the research has employed some form of highly specialized materials such as simple lists of words or words embedded in repeated carrier sentences. Until the more recent history of speech study, those types of specialized materials were fruitful enough in the field that little work was done toward assessing what the speech signal and its theorized sounds and cues look like in more ordinary circumstances, except to some extent in the engineering field, where researchers were interested in transmitting speech efficiently via electronic apparatus.

Recently, however, there has been a flurry of interest in spontaneous speech and speech reduction, as researchers find that the kind of everyday speech people are engaged in is quite different from the carefully articulated “laboratory speech” the field has been eliciting, studying, and theorizing about for so long.

One of the main findings to come from this recent activity in spontaneous speech is that the acoustic cues thought to define speech sounds under the laboratory speech understanding are often reduced, missing, or take an unexpected form (Johnson, 2004; Cheng and Xu, 2009; Ernestus and Warner, 2011; Warner, 2011), yet this makes little trouble for listeners. The ability of listeners to adapt to the huge variety of realizations of speech sounds in casual, spontaneous, conversational speech is remarkable, and is deserving of the attention that is now developing.

One type of speech sound whose conversational realizations have not yet been studied in depth is lexical tone. In tone languages, apart from the consonants and vowels making up the syllables and words, a pattern of vocal pitch, much like singing, is also part of the specification of the words, such that producing the wrong pitch pattern during a syllable may count as saying the wrong word, or in any case counts as a mispronunciation. As an example that many a learner of Mandarin Chinese has chuckled about, a common way of politely catching the attention of a stranger to ask a question or directions begins, ⟨请问…⟩ qǐngwèn 'May I ask…', with a dipping pitch in the first syllable, and a falling pitch in the second. However, with another dipping tone in the second syllable in place of the falling pitch, the phrase becomes ⟨请吻…⟩ qǐngwěn 'Please kiss…'. In tone languages these pitch patterns serve much the same purpose as segments, and are part of the specification of words. However, because of their pitch-centered nature, tones are also quite different from segments in some ways (Taft and Chen, 1992; Cutler and Chen, 1997), and interact with many other features of speech such as intonation, stress, and focus (Xu, 2004). The contribution of tone in the Mandarin phonological system has been quantified (Surendran and Levow, 2003), and determined to carry as much information in the lexicon as vowels do. Given its central importance in the world's most spoken language (Lewis et al., 2015), and the commonplace occurrence of tone languages around the world (Dryer and Haspelmath, 2013), a study of Mandarin tones and their manifestation in everyday conversation seems appropriate.

This chapter reports on a study of native perception of Mandarin tones in hummed and resynthesized words recorded in conversation and contrasting with words recorded in a careful word list reading, in order to determine how much information about tone categories is available to listeners in the pitch contour, the segmental acoustic detail, and the whole signal, and how these sources differ in the two speech styles. One aspect of what might be regarded loosely as "normal conditions" (among many, of course) is the acoustic reality of the word forms produced in everyday casual conversation. It is known, for example, that at least for many languages, the acoustic information we use to perceive segments is less robust in conversation than in careful speech (Ernestus and Warner, 2011; Warner, 2011; Warner and Tucker, 2011), so that cues that characterize a sound in careful speech (the sort typically studied in the laboratory) may be reduced or absent in conversation. Whole segments, syllables, and even words have been found missing any detectable acoustic traces in the speech (Johnson, 2004). These same types of reduction have been reported in Taiwan Mandarin (Ernestus and Warner, 2011; Tseng, 2004a; Cheng and Xu, 2009; Cheng et al., 2010, 2011; Cheng, 2012; Cheng and Xu, 2013, 2014), and one study found a reduced tone space in repeated mentions of target words in mainland Mandarin during a map task (Berry, 2009), which perhaps approximates conversation. If tones are reduced in conversation, how does that affect listeners' use of the available information sources, and how does that differ from the case of careful speech?

Mandarin is an excellent language for the study of lexical tone. In addition to native speakers being readily found, the tone system is a near minimal set of all the various properties that make tone systems interesting (Fromkin, 1978; Yip, 2002). It has level and contour tones, full and "neutral" tones, stress, tone Sandhi, voice quality cues, and so forth. And all of this occurs in a system of 4 full tones, and the neutral tone, a distinction which itself merits interest. The tones have come to be identified by native speakers as simply "tones 1-4" (or sometimes the "high tone", "rising tone", "dipping tone", and "falling tone" respectively), and "the neutral tone" or "clear tone". In Pinyin romanization, the tones are designated by diacritics (¯, ´, ˇ, `, and no diacritic for the neutral tone) above the syllable nucleus letter, where the diacritics are iconic of the pitch pattern to be produced during a syllable of that tone. The tones are predominantly distinguished from each other by the pitch contour produced over the voiced portion of a syllable (Chao, 1965; Duanmu, 2007; Lin, 2007).

Schematic pitch contours for each of the four full tones are shown in Figure 2.1.

Tones 1 and 4 are readily distinguished from the other tones and each other (even when the pitch space is artificially compressed to a range of just 2Hz (Klatt, 1973)), while tones 2 and 3 are frequently confused (Blicher et al., 1990; Shen and Lin, 1991).

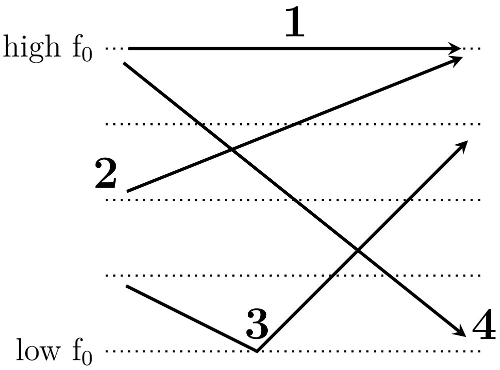


Figure 2.1: Schematic of the standard pitch patterns of tones 1–4 as conceptualized by native speakers, and taught to children and L2 learners. The neutral tone (not shown) meanwhile, is relatively short and varies widely in its pitch realization according to its neighboring tones (Cao, 1992; Chen and Xu, 2006). Horizontal dotted lines represent a 5-level division of a speaker's normal speaking pitch range. This visual scheme was the basis of the current IPA standard notation for tones, devised by Chao (1930). This figure is repeated for reference.

Mandarin tone 3 syllables undergo Sandhi (Shih, 1997; Chen, 2000; Cheng et al., 2013), phonological processes which alter the usual pitch contours of some tones in specified environments. The most prominent form in Mandarin is a process which transforms the dipping tone (tone 3) into a rising tone (tone 2) before another dipping tone. This process is cyclical, such that a run of tone 3s will become tone 2s, except for the final one which stays a 3. Prior to any other tone in a word, tone 3 is realized as a "half-tone-3" (Chao, 1965), which is just the low falling portion of the full tone 3 pitch contour. There has been some debate whether tone 3s in the Sandhi environment are truly indistinguishable from tone 2s. Cheng et al. (2013), for example, find acoustic differences in pitch velocity between underlying tone 2s and Sandhi tone 3s, raising the question of whether listeners are able to detect these differences, and if so, under what conditions. Most accounts posit that the two derivations of an ostensible tone 2 pitch pattern are equivalent, and barring further information to indicate that it is indeed an underlying tone 3, listeners should associate the pattern with tone 2.

Mandarin also utilizes an unstressed "neutral" tone (sometimes shown as "tone 0" to follow.), which is described as a "weakly articulated" tone, meaning its realizations are highly dependent on the syllable's surrounding tones. According to Chao (1965), the neutral tone appears after a tone 1 as half-low, after tone 2 as mid, after tone 3 as half-high, and after tone 4 as low. This means that listeners should in theory be able to ascertain the prior tone from the form of a following neutral tone, which may become important for the present study. Listeners are predicted to do well in identifying first syllable tones followed by a neutral tone, and this performance may be robust to the manipulation of this first syllable itself.

In addition to vocal pitch trajectories, the duration of syllables (Blicher et al., 1990; Chang and Yao, 2007; Cabrera et al., 2014), the amplitude contour (Whalen and Xu, 1992; Liu and Samuel, 2004), voice quality (Yu, 2010; Bissiri et al., 2014), and segmental variations (Xu, 1997; Hu, 2004) all provide information about tone categories, just as segmental categories can be signalled by multiple correlated cues, e.g. duration, F_1 onset lag, and pitch at onset of voicing can all signal voicing contrasts (Schertz, 2014).

A number of studies have used pure tone or resynthesized stimuli to study cue combinations in the perception of tones (Abramson, 1972; Howie, 1976; Whalen and Xu, 1992; Liu and Samuel, 2004, and others), but none have extended the analysis of information sources into conversational word forms. In fact, very few studies of any kind consider conversational speech in Mandarin in the phonetics domain (Berry, 2009; Tseng, 2004b, are exceptions).

2.2 Methods

2.2.1 Materials

One[1] 28-year-old male Beijing-area native Mandarin speaker was recorded in a 20-minute telephone conversation[2] with a close Beijing-area male friend (not present on site) in a noise-attenuating recording booth at the Chinese Academy of Social Sciences in Beijing in the summer of 2013. This was free conversation; no topics were offered. Only the speaker in the lab was consented and recorded, and his interlocutor was informed of this at the beginning of the conversation.

The 44.1kHz/16-bit WAV recording was made via a Countryman E6 omnidirectional head-mounted microphone (on the opposite ear from the telephone) with 0dB flat cap, a Shure FP23 preamp unit, and a Sony PCM-M10 recorder. This setup was chosen for its flat and faithful frequency response between 20Hz and 20kHz.

[1] Because of the many factors in the design of the present study, and limitations on the running time of the experiment for each listener (as it is, the experiment takes roughly forty minutes to a full hour, a very long time to sit and classify the tones of isolated words), only one speaker was used in the preparation of the stimuli. This was a choice of trade-off between generalization of speaker and generalization of the other factors.

[2] Although there are differences between conversing on the telephone and conversing face-to-face (Moreno and Stern, 1994; Reynolds et al., 1995, report signal processing consequences of this difference), telephone conversation is a familiar, natural task, there were no obvious instances of Lombard reflex (Zhao and Jurafsky, 2009), yelling, or difficulty communicating, and the content of the recorded speech—the illness of a friend's child; candid observations of the temperament of classmates; the arranging of daily schedules with a partner; etc.—is indicative that the conversations recorded are casual.

From this recording, 5 words each for all of the 20 possible tone combinations for two-syllable words (4 full tones are possible in the first syllable, and all 4 plus the neutral tone are possible in the second syllable) were selected at random using a script which positions a Praat (Boersma and Weenink, 2015) editor window at a random time point within the length of the recording. The closest two-syllable word to that time-point was selected, and then the editor was again placed at a random time-point until each of the 20 tone combinations had 5 different representative words in the recording. Any two-syllable words were used, and in a few cases, when appropriate two-syllable words for a given tone combination were not forthcoming, a simple two-syllable phrase was used instead, for example ⟨很好⟩ hěn hǎo 'very good' or ⟨好多⟩ hǎo duō 'really a lot'. These 100 words were then randomized and compiled in a list, and 3 additional filler words were added to the beginning and 3 to the end of the list to eliminate inconsistent intonation effects at the ends of the list while reading. After several days, the speaker was then again invited to the lab to record a reading of the word list. This word list was recorded in the same recording booth with the same settings on the same equipment as the conversation had been. Each word in the list was read three times. The speaker subsequently reported that he was not aware that the words had been selected from his conversational recording.

One token from each of the 100 conversational words and 100 read words were identified for use in creating the stimuli for this project. In the word list recording, the first token of the 3 repetitions of each word was selected unless it had some kind of intrusive acoustic defect such as noise or clipping. Due to a minor listing error, one word list entry, ⟨如果⟩ rúguǒ ‘if’, had no counterpart in the conversational recordings, so another tone 2-tone 3 word with similar syllable structures, ⟨没有⟩ méiyǒu ‘has not’ was identified in the conversational recording to serve as its counterpart. Time-aligned syllable-level labels were created by hand for all 200 words.

For each of the 200 words selected, two types of acoustic manipulation were applied – synthetic humming of an extracted pitch track, and Linear Predictive Coding (O’Shaughnessy, 1988) resynthesis – to yield 3 stimulus versions. Synthetic humming was performed in Praat (Boersma and Weenink, 2015) by extracting the vocal pitch (Boersma, 1993) and then applying a stylized synthetic hum model to the pitch track to eliminate segmental detail and voice quality cues, providing only pitch information, modulated with the original intensity envelope of the signal.

Resynthesis was performed via Praat’s (Boersma and Weenink, 2015) LPC functionality, with a maximum frequency of 5000Hz, 20 pole spectral matching, 10ms step, 25ms window length, and 6dB/octave pre-emphasis beginning at 50Hz. LPC models the contributions of the glottal source (the voicing mechanism) and the supralaryngeal vocal tract filter (which forms the resonating chamber), allowing independent manipulation of the contributions of these two factors to the output sound to create stimuli that are otherwise identical, but in one case retain pitch information but no segmental information, while in another the segmental information is preserved, but the pitch information is removed. In the present case, this separability will allow the removal of pitch information by recreating the speech without any glottal pulsing. The three stimulus versions differed only in the humming or resynthesis of the first syllable, which is the syllable whose tone category listeners will identify; the second syllable in all words in all resynthesis conditions were fully resynthesized, i.e. LPC separated the contributions of source and filter, and the resulting models were used to regenerate the sound. All stimuli were resynthesized similarly for consistency of unnaturalness due to the resynthesis.

In the "full" stimuli, the first syllable was also simply fully resynthesized in this manner, so that both vocal pitch information and segmental information are present.

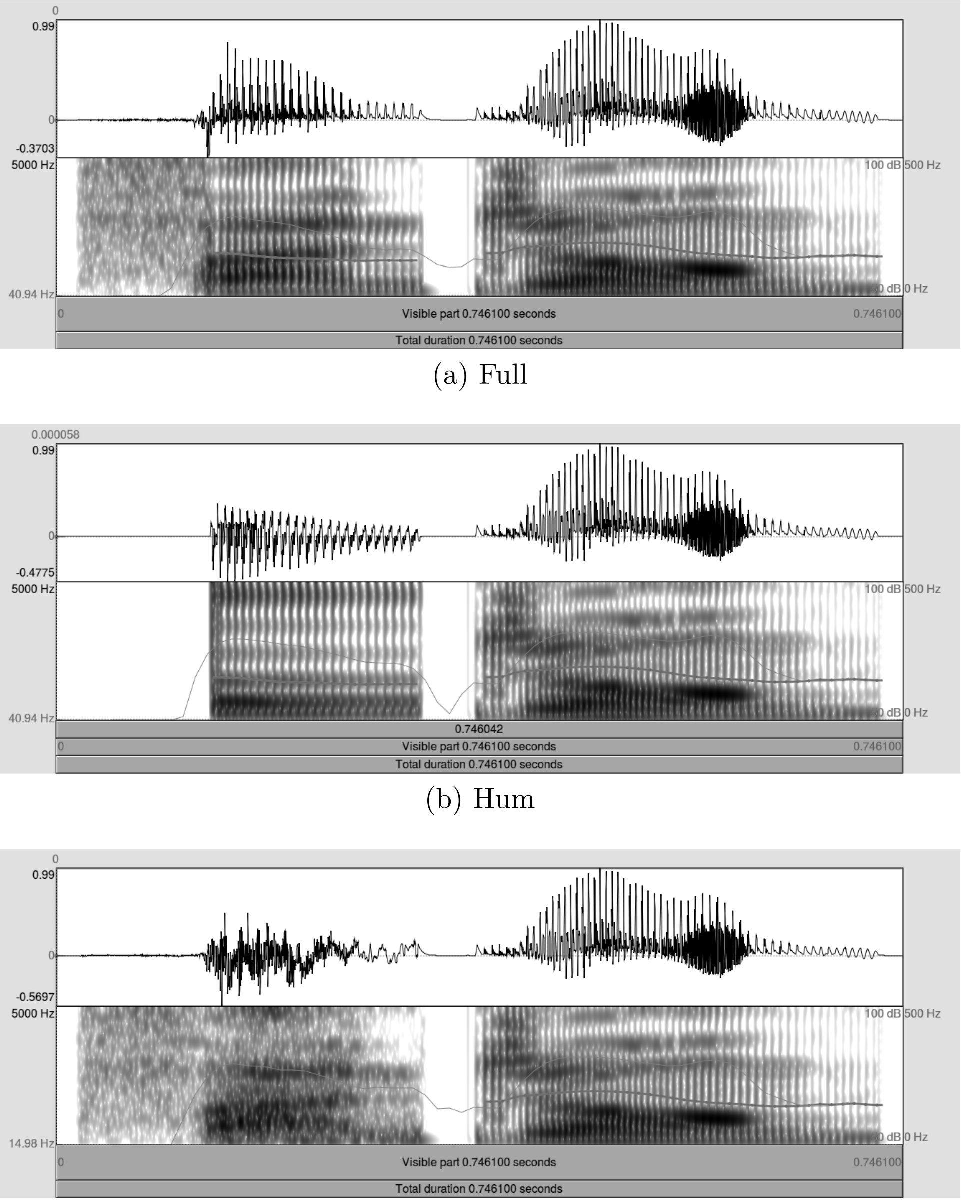
This condition serves as a reference from which to compare the remaining resynthesis conditions, and once again was resynthesized for comparability so that differences between “full” and the other conditions would not be an artifact of whether or not they were manipulated.

In the “hum” stimuli, the pitch track of the first syllable was hummed as mentioned above, using a simple voicing model provided by Praat. This voicing provides pitch information in a human-like voice, without segmental or voice quality information. The original intensity curve was then reapplied to the phonated pitch curve, so that the pitch information available to listeners more closely mimicked the original signal pitch information, and to avoid any unnatural jarring artifacts during consonants. These stimuli appear to listeners to have a hummed first syllable with the same pitch and intensity of the original, followed by the full second syllable with complete information.

For “whisper” stimuli, conversely, the vocal source pitch pulses were removed, and the syllable was resynthesized to create a whispered version of the first syllable. This process is similar to other studies on whispered speech in tone languages, such as Abramson (1972), Howie (1976), or Liu and Samuel (2004). Once again the original intensity profile is restored, creating a whispered version of the first syllable without the original pitch contour. Although Howie's study for Mandarin showed poor tone discrimination of vocoded pseudo-whisper, in Abramson's Thai data, listeners performed better with similar pseudo-whisper stimuli than with naturally whispered stimuli, prompting Whalen and Xu (1992) to suggest that listeners may nonetheless glean some information about tonal categories from formants or vowel quality changes preserved in the resynthesized stimuli. It is this segmental information, of which intensity is one sort, which is captured here in the present "whisper" resynthesis. Figure 2.2 shows sample stimuli from the three resynthesis conditions.

2.2.2 Procedure

All of these resulting stimuli were presented using studio quality over-ear headphones using E-Prime (Schneider et al., 2012), to 20 native Mandarin students (aged 18-29, mean 23; all grew up in China, and have been living in the U.S. for between 1 and 7 years) at the University of Arizona in Tucson. Participants were monetarily compensated for their participation. The task was to identify the tone category of the first syllable by pressing one of four buttons on a button box. The buttons were



(a) Full

(b) Hum

(c) Whisper

Figure 2.2: Waveforms, spectrograms, pitch (dark line) and intensity tracks (thin grey line) of perception experiment stimulus examples from the three Resynthesis conditions for the word ⟨反正⟩ fǎnzhèng ‘anyway; in any case’.

clearly labeled 1 – 4 by a layout on the computer screen showing the buttons and indicating tones 1 – 4 with numbers and also the iconic Pinyin tone diacritics (¯, ´, ˇ, `) that Mandarin speakers are accustomed to associating with the tones in dictionaries, teaching materials, and so forth. Participants were asked to keep the forefinger and middle finder of both hands on the four buttons at all times during the experiment to facilitate quick responses, and to respond as quickly and accurately as possible.

During the experiment, all instructions on screen were in standard Mandarin text, composed with the help of a native Mandarin speaking colleague. Because the task required is not a familiar one, and the details are important, each subject received oral instructions in Mandarin prior to beginning the experiment, and then were presented with text instructions at the start of the experiment run. Instructions were detailed and contained clear examples, e.g. "If you hear the word ⟨海边⟩ (hǎibiān; hai3bian1), you would press button 3, because the first syllable is ⟨海⟩ which has the 3rd tone. … Below are some further examples of 2-syllable words, and the tones of their 1st syllables. …"[3] The instructions also contained a description of the blocking of the

[3] This example is translated. All text on-screen was in standard Mandarin characters.

experiment, explaining that in the first part, they would hear whole words, in the second part, they would hear words with whispered 1st syllables, and in the last part, they would hear words with hummed 1st syllables. They were also told that they would be listening to some words recorded from word lists, and some recorded from actual conversations. Each resynthesis condition block contained half wordlist recordings, followed by half conversational recordings, and the practice items were organized this way also.

The experiment began with 12 practice items (4 full, 4 whisper, 4 hum; half word list, half conversation), then presented test items in “full”, “hummed”, and “whispered” blocks (as detailed above in §2.2.1), with breaks in between, and text pages explaining which kind of stimuli were to be presented in the coming block. Within each block, the 100 word list recorded items were presented, followed by the 100 conversational items. All subjects received all items in all conditions.

In this experiment, each item is presented in each condition, producing the possibility of lingering activation (Forster and Davis, 1984) at subsequent presentations. However, it is unlikely that frequent words such as those chosen for stimuli would facilitate the selection of a first syllable tonal category as is the task in the present experiment, or affect their error rates in doing so.

The blocking was chosen to capitalize on the direct comparability of the speaker's conversational and word list items, but also to give listeners a good arrangement for acclimatization to the task. A pilot study using the reverse blocking (first conversational whisper, then word list whisper, conversational hum, word list hum, conversational full, and word list full) made clear that even with an initial practice block to familiarize the listeners to the task, when the experimental blocks began with whispered conversational stimuli, the listeners were extremely confused by the very short conversational items, and responses were not indicative of the perception performance intended. Thus, blocking was chosen to begin with full word list items so that the beginning of the experiment would be reasonable for settling into the task required, and later conversational and less informative items would occur when the task was clear.

2.3 Results

Due to a copying error, the first 10 subjects received items in the wrong block during part of the experiment. They were excluded, the experiment corrected, and a further

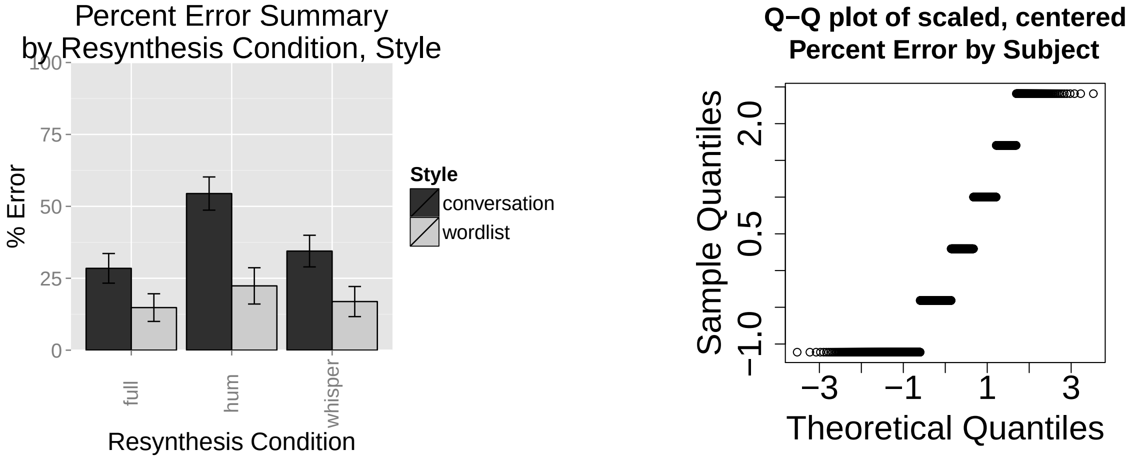
20 subjects participated. These latter 20 subjects' results are reported here.

The dependent variables were responses and response latencies. What follows are general summary distributional facts about the collected samples, ANOVA accuracy results, and an assessment of individual differences among listeners.

2.3.1 General data summary, percent error

A statistical analysis follows below. First, Figures 2.3 and 2.4 present summary distributions of accuracy by subjects. These are plotted as percent error to facilitate comparison with response time data. Tones of word list stimuli were better identified than those in conversational tokens. The effect of resynthesis condition is larger in the conversational stimuli than in the word list stimuli. In both styles, tones in the hum condition are the most poorly identified, with tones in whisper stimuli nearly as identifiable as in the full signal. There is an interaction of Style and Resynthesis condition, with Resynthesis affecting perception accuracy in all conversational tone contexts, but more limited contexts in the word list items.

Conversation clearly has flatter distributions of responses for each tone, showing higher error rates than in the word list items. This was expected due to the reduced tone space (Berry, 2009) and high degree of tone cue reduction as reported in Brenner (2013). Each style has roughly equal accuracy for all four tones. In the wordlist, accuracy is quite stable across resynthesis treatments, while identification of the conversational tones are markedly poor in the hummed condition. Section 2.3.3 gives a summary of common tone confusions in the two styles.



(a) Percent error by resynthesis condition and speech style.

(b) Q-Q plot of scaled mean percent error scores. This compares data quantiles (Y-axis) to normal distribution quantiles (X-axis).

Figure 2.3: % Error distribution summary and Q-Q plot.

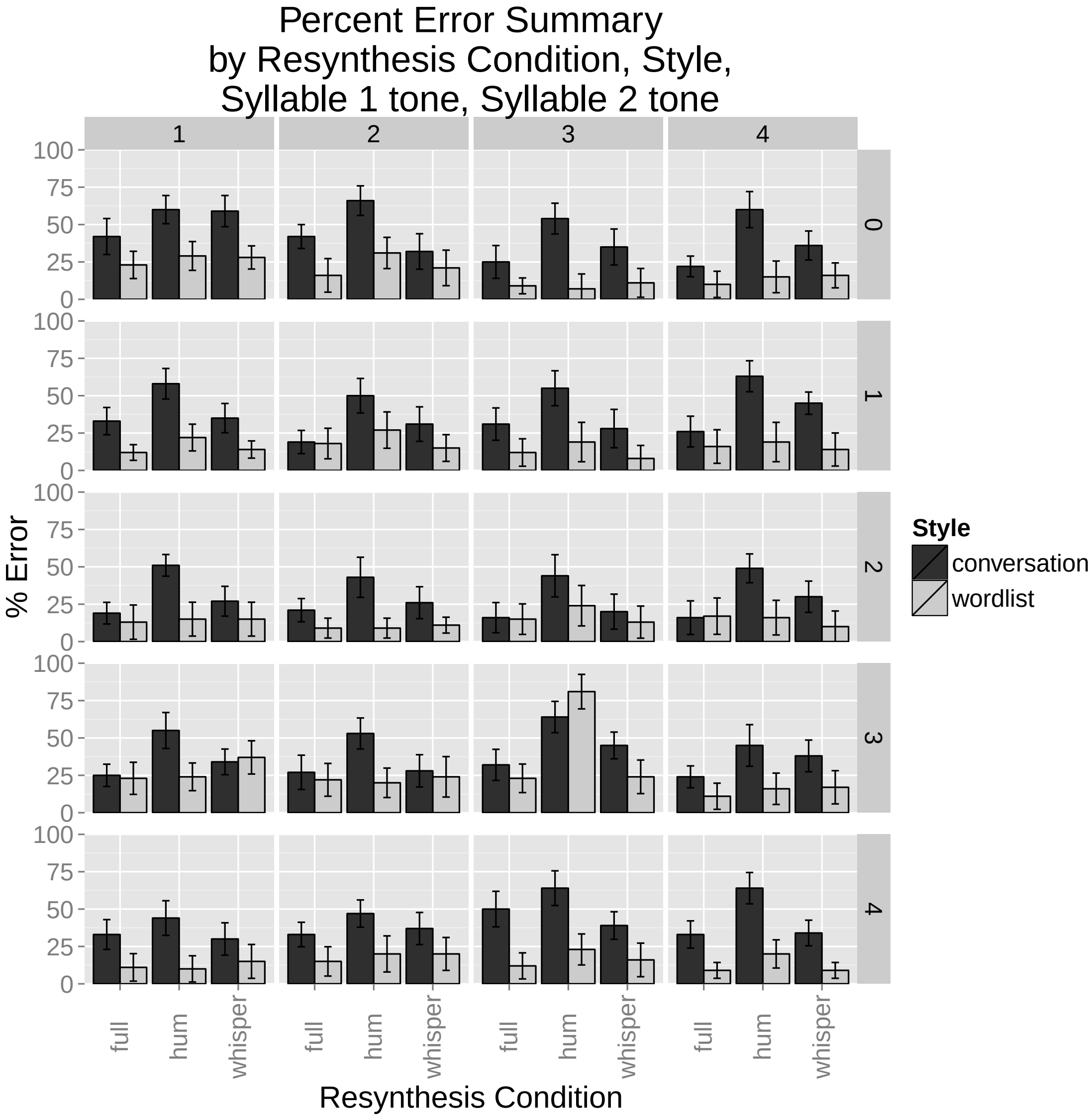


Figure 2.4: Plot grid columns are 1st syllable tones, rows give 2nd syllable tone context. Error bars indicate 95% confidence intervals.

2.3.2 Error rate ANOVA comparisons

A 2 × 3 × 4 × 5 repeated measures by-subjects (F1) / by-items (F2) ANOVA[4] was computed over the raw percent error scores. As is shown in Figure 2.3b, the percent error scores lie well-clustered close to the diagonal normal line, indicating near normal distribution. For this reason, no transformation was applied. This was a fully within-subjects design, with factors Style ("conversation", "word list"; the speech style of the stimulus), Resynthesis condition ("full", "hum", "whisper"; the type of resynthesis applied to the first syllable), First syllable tone (1-4; the tone to which listeners responded), Second syllable tone (0-4; the tonal context. "Tone 0" refers to the neutral tone), and all interactions of these four factors. In the by-items analysis, the factors Style and Resynthesis condition are within-items, while First syllable tone and Second syllable tone are between-items factors.

The omnibus ANOVA model revealed significant effects of Style ($F1_{(1,19)} = 309.4$, $p < 0.001$, $F2_{(1,79)} = 106.89$, $p < 0.001$), Resynthesis condition ($F1_{(2,38)} = 9.0$, $p < 0.001$, $F2_{(2,158)} = 87.36$, $p < 0.001$), and Style × Resynthesis condition ($F_{(1,2)} =$

[4] Although a logistic linear mixed effects (LME) model would provide simultaneous control for the crossed random effects of speakers and items (Lindstrom and Bates, 1988; Barr et al., 2013), convergence failures and lengthy computer runtime fitting the models in the current study's large parameter space prevented the practical application of LME in this case.

38, $p < 17.40.001$, $F2_{(2,158)} = 31.32$, $p < 0.001$). The effects of First syllable tone ($F1_{(3,57)} = 1.0$, $p < 0.5$, $F2_{(3,79)} = 0.54$, $p < 1$), Second syllable tone ($F_{(4,76)} = 29.8$, $p < 0.001$, $F2_{(5,79)} = 2.31$, $p < 0.1$), First syllable tone × Second syllable tone ($F1_{(12,228)} = 8.5$, $p < 0.001$, $F2_{(12,79)} = 1.43$, $p < 1$), Style × First syllable tone ($F1_{(3,57)} = 2.3$, $p < 0.1$, $F2_{(3,79)} = 0.41$, $p < 1$) Style × Second syllable tone ($F1_{(4,76)} = 24.9$, $p < 0.001$, $F2_{(5,79)} = 1.65$, $p < 1$), Resynthesis condition × First syllable tone ($F1_{(6,114)} = 2.4$, $p < 0.05$, $F2_{(6,158)} = 2.12$, $p < 0.1$), Resynthesis condition × Second syllable tone ($F1_{(8,152)} = 2.6$, $p < 0.01$, $F2_{(10,158)} = 0.84$, $p < 1$), Style × First syllable tone × Second syllable tone ($F1_{(12,228)} = 4.2$, $p < 0.001$, $F2_{(12,79)} = 0.40$, $p < 1$), Style × Resynthesis condition × First syllable tone ($F1_{(6,114)} = 2.8$, $p < 0.05$, $F2_{(6,158)} = 1.63$, $p < 1$), Style × Resynthesis condition × Second syllable tone ($F1_{(8,152)} = 2.5$, $p < 0.05$, $F2_{(10,158)} = 0.93$, $p < 1$), Resynthesis condition × First syllable tone × Second syllable tone ($F1_{(24,456)} = 2.6$, $p < 0.001$, $F2_{(24,158)} = 1.49$, $p < 0.1$), and Style × Resynthesis condition × First syllable tone × Second syllable tone ($F1_{(24,456)} = 3.4$, $p < 0.001$, $F2_{(24,158)} = 1.45$, $p < 0.1$). were not significant.

Now collapsing over First syllable tone and Second syllable tone (since they did not participate in any main or interaction effects in the by-items analysis, likely due to low power in these between-items factors), the data were split by speech Style due to the interaction with Resynthesis condition. In the word list items, the effect of Resynthesis condition was not significant by the by-subjects analysis ($F1_{(2,38)} = 1.15$, $p < 1$, $F2_{(2,198)} = 11.68$, $p < 0.001$), but in conversational items, it was significant by both analyses ($F1_{(2,38)} = 35.62$, $p < 0.001$, $F2_{(2,198)} = 83.53$, $p < 0.001$), with "hum" error rates significantly higher than "whisper" ($F1_{(1,19)} = 37.17$, $p < 0.001$, $F2_{(1,99)} = 75.00$, $p < 0.001$), but "whisper" error rates not significantly different from "full" by the by-subjects analysis ($F1_{(1,19)} = 2.00$, $p < 1$, $F2_{(1,99)} = 11.78$, $p < 0.001$).

The individual tones did not participate in any effects which were significant by the by-items analysis, but there is strong theoretical reason to expect that identification performance should not be uniform across tone category combinations, as detailed in §2.1, since a neutral tone in the second syllable is theorized to be informative of first syllable tone categories, and tone 3 is expected to be perceived poorly in the Sandhi environment (preceding another tone 3). It may be that the many tone categories prevent the detection of these differences which occur in only limited levels of the tone factors. For this reason, an exploratory by-subjects analysis retaining the tone contexts, and splitting by Second syllable tone, was pursued, and is summarized in

Table 2.1.

Wordlist	F1/F2	F1/F2	F1/F2
(dfs)	(3,57)/(3,16)	(2,38)/(2,32)	(6,114)/(6,32)
Syll 2 tone	S1 tone	Resynth	S1tone × Resynth
0	13.94***/1.39	0.72/2.12	1.45/0.80
1	1.85/0.36	1.70/4.67*	0.37/0.23
2	2.55/1.72	0.50/1.68	1.79/1.41
3	10.45***/5.02*	4.59*/11.59***	14.20***/12.55***
4	0.98/1.70	0.39/2.50	1.48/0.88
Conversation	F1/F2	F1/F2	F1/F2
(dfs)	(3,57)/(3,16)	(2,38)/(2,32)	(6,114)/(6,32)
Syll 2 tone	S1 tone	Resynth	S1 tone × Resynth
0	6.22**/0.86	21.29***/25.02***	2.45*/0.07
1	2.45/0.30	35.57***/27.04***	1.03/0.74
2	0.92/0.16	21.98***/14.49***	0.21/0.16
3	2.23/0.30	21.49***/12.18***	0.79/0.34
4	4.15**/0.58	10.49***/8.96***	0.96/0.68

Table 2.1: Percent error summaries by Style and Second syllable tone, for First syllable tone, Resynthesis condition, and First syllable tone × Resynthesis condition. Values shown are F-ratios. *$p < .05$; **$p < .01$; ***$p < .001$

For all Second syllable tones in conversation, there was a significant effect of Resynthesis condition, with Welch two-sample t-tests revealing that "hummed" items have much higher error rates than "whispered" items in each case (all $p < 0.05$ for both by-subjects and by-items analyses), but non-significant differences between "whispered" and "full" items (always $p > 0.05$ in both analyses, except for a significant by-subjects difference only for tone 3 ($p < 0.05$)).

In the word list items with Second syllable tone 3 (the Sandhi context), First syllable tone interacts significantly with Resynthesis condition, prompting further division of these stimuli by First syllable tone. The results are summarized in Table 2.2. As the table shows, the effect is driven by First syllable tone 3, in the Sandhi environment. In this environment, once again “hum” has significantly higher error rates (in excess of 80%) than “whisper” (averaging roughly 25%), but there is no significant difference between “whisper” and “full” (around 30%).

Wordlist	F1/F2
(dfs)	(2,38)/(2,8)
Syll 1 tone	Resynth
1	3.59*/1.58
2	0.14/0.13
3	54.78***/99.46***
4	0.13/0.34

Table 2.2: Second syllable tone 3, word list items: Percent error summaries by First syllable tone and Resynthesis condition. Values shown are F-ratios. *$p < .05$; **$p < .01$; ***$p < .001$

Having detailed the patterns of statistical effects in error rates, it is useful to consider which mistakes listeners made. Confusion matrices will provide fuller detail in the individual kinds of responses that were made.

2.3.3 Confusion matrices

Next, Tables 2.3a-2.3h give confusion matrices for each tone, resynthesis condition, speech style. Confusion matrices tell us not only how often errors occurred, but which errors occurred, and can indicate how symmetrical tone confusions are, and where errors may reflect perception bias. Categorization accuracy is reflected in proportions on the diagonal (e.g. tone 2-response 2, tone 4-response 4, etc.), while inaccurate categorization shows as off-diagonal proportions (tone 3 responded to as tone 2). Symmetrical confusion is reflected in symmetrical values across the diagonal (e.g. tone 4-response 1, tone 1-response 4), and perception bias shows as asymmetrical values across the diagonal. Statistical testing of the percent error patterns appears in §2.3.2.

Conversational tone 1 in the full condition is often mistaken for tone 4, but word list tone 1 was not.

Tone 3 in the hum conditions of both speech styles was often misclassified as tone 2 due to tone Sandhi and a lack of any segmental detail to disambiguate the merger. Recall that a tone 3 followed by another tone 3 surfaces as a tone 2 (with some caveats; discussion of Mandarin tone Sandhi appears in §2.1).

		Response			
		1	2	3	4
Tone	1	77	9	3	7
	2	5	79	10	3
	3	3	14	78	3
	4	5	4	5	83

(a) Total, word list

		Response			
		1	2	3	4
Tone	1	56	17	9	14
	2	14	60	13	10
	3	8	19	58	13
	4	14	15	9	58

(b) Total, conversation

		Response			
		1	2	3	4
Tone	1	80	7	3	6
	2	5	81	10	2
	3	3	9	84	2
	4	5	3	5	83

(c) Full, word list

		Response			
		1	2	3	4
Tone	1	66	9	4	17
	2	12	67	9	8
	3	7	15	66	8
	4	10	9	5	70

(d) Full, conversation

		Response			
		1	2	3	4
Tone	1	77	9	2	9
	2	6	77	11	5
	3	3	25	66	3
	4	6	6	6	80

(e) Hum, word list

		Response			
		1	2	3	4
Tone	1	45	26	13	15
	2	19	45	19	14
	3	9	28	43	19
	4	23	20	14	42

(f) Hum, conversation

		Response			
		1	2	3	4
Tone	1	73	10	5	7
	2	4	79	11	4
	3	2	8	84	4
	4	4	4	5	84

(g) Whisper, word list

		Response			
		1	2	3	4
Tone	1	59	17	9	11
	2	12	67	11	8
	3	8	15	64	11
	4	11	17	9	61

(h) Whisper, conversation

Table 2.3: Confusion matrices: percent total response distributions for each tone as a whole, and for each of the three resynthesis conditions in the word list items and in the conversation items. Due to timeout non-responses, rows may not sum to 100%.

2.4 Discussion

The ANOVA analyses in §2.3 give a complex picture, but that was anticipated, since a 4-way interaction was likely from the design. The lack of more significant interactions is likely due to the many levels of the tone factors.

In the careful speech word list items, listeners performed remarkably consistently in nearly all resynthesis conditions, for nearly all tone combinations, averaging 21% error overall. There was one significant deviation from this consistency.

Tone 3 initial syllables in words with a second syllable tone 3 were badly misclassified (83% error) in the "hummed" items due to insufficient lexical information to infer Sandhi had taken place even with the original intensity contours preserved, since tone 3 before another 3 is near indistinguishable from a tone 2 (Wang and Li, 1967), though Cheng et al. (2013) do find an acoustic difference in tone velocity contours between the Sandhi tone 3 and tone 2, but it is unclear whether listeners can make use of that cue effectively. Although conversational tone 3 items in general are poorly identified, in the Sandhi context they were better identified than the careful speech items (64% error; Welch two-sample t-test, t=2.40, df=37.7, p=0.022), suggesting that prosodic cues in the pitch track or amplitude envelope of the conversational items are providing more information about the identity of the underlying tone in this Sandhi context (Shih, 1997, provides a summary of literature and phonological theory on the role of prosody in Mandarin tone Sandhi).

There was no significant effect of the predicted second syllable neutral tone on the error rates of different first syllable tones, but the error rates are suggestive, with twice as many errors for tones 1 and 2 than for 3 and 4. Neutral tone syllables are known to tend to assimilate in their pitch contours to surrounding tones (Cao, 1992; Chen and Xu, 2006), but their realizations also provide information about the syllables that precede them. These data suggest that the preceding tones may be influenced by the neutral tone, with the tones ending on high pitch affected the most, but a more targeted study is needed.

In conversational items, there were significant effects of resynthesis condition throughout, in all cases with "hummed" items faring poorly for tone identification, and "whisper" items patterning with "full" items. The impoverished acoustic signal in the "hummed" stimuli were insufficient in the conversational context where items are shorter and often have reduced acoustic cues. The "whispered" items appear to retain enough information in the segmental, prosodic, and perhaps lexical information to identify the tones.

Overall, the word list tones are more reliably and consistently identified regardless of resynthesis or tone combinations. Conversational tone perception is more contingent on the available information sources and tone combinations involved. With twice the proportion of errors, conversation is clearly a more complex, more challenging listening context for tones as it has been shown for segments, and “whisper” provides more of the needed information to tone than does the pitch contour preserved in “hum” items.

Chapter 3

Pitch and Tones In Vivo

3.1 Introduction

This chapter[1] describes an experiment constructed with genuine conversational stimuli to assess the actual contribution of pitch to Mandarin listeners' comprehension of ordinary everyday speech. Vocal pitch appears to be utilized in all spoken languages[2] to signal phrase- and larger- level meanings, such as attitude or emotional stance, emphasis or contrast, declarative, imperative, or interrogative utterances, and these kinds of use of vocal pitch are lumped together under the term "intonation". In nearly half of the world's languages (Maddieson, 2011), however, vocal pitch patterns on a smaller scale—the syllable or word— are used to differentiate those smaller units, so that, for example, in Burmese /ma/ with a low pitch conveys the word 'hard', while /ma/ with a high pitch signals 'towering' (Watkins, 2001). This difference in pitch in

[1] Some of the introductory material is common to the other chapters, and is described here as it is relevant to this particular study.

[2] The exceptions indirectly implied here are signed languages such as American Sign Language, Chinese Sign language, Turkish Sign Language, and many others (Zeshan, 2008; Premaratne, 2014).

the two syllables is so significant in the Burmese sound system that it differentiates words just as consonants and vowels do, and these pitch patterns associated with the identity of words are called “tones”.

Tone pitch contours are overlaid on the intonation pitch contours like “small ripples riding on large waves” (Chao, 1965, p.39). In fact, the pitch contour resulting in speech is a composite shape arising from the combination of tone, intonation, emphasis, focus, attitudinal characteristics, etc., making it a highly simplified output of a very complex set of inputs (Chao, 1933; Xu, 2004). It may be that this density of information realized in a simple curve is too ambiguous for listeners to parse out all of the contributing factors, making other streams of information in the signal preferable or more reliable for some purposes.

Because of the vital contribution of pitch in the system of sounds used to signal words in tone languages, it might seem obvious that listeners must rely on pitch in the spoken acoustic signal to help them in comprehension, but a number of studies with naturally or artificially whispered speech in tone languages suggests that listeners are able to understand this whispered speech without much trouble (Howie, 1976; Abramson, 1972; Liu and Samuel, 2004; Chang and Yao, 2007). These existing studies have used isolated words or careful speech in their stimuli, so the question remains how much, under everyday listening conditions, listeners rely on pitch information in comprehension. Therefore, the study reported here utilizes artificially whispered conversational utterances to compare comprehension error rates with and without pitch information.

As demonstrated in a growing number of studies on speech reduction and conversational or spontaneous speech (Berry, 2009; Johnson, 2004; Cheng and Xu, 2009; Ernestus and Warner, 2011; Warner, 2011), many of the individual acoustic cues identifying speech sounds, including tones, are frequently reduced or absent in conversational speech. This raises the question of whether categories such as tone which are perceived more slowly and with more errors (Cutler and Chen, 1997) simply become irrelevant in the fast-moving context of conversation. Do Mandarin listeners really make use of tone in their conversational comprehension, and if so, how much? And do they get information about tone categories directly from the pitch patterns, or can they just as well extract the category identity from other cues? A number of studies have used pure tone or resynthesized stimuli to study cue combinations in the perception of tones (Abramson, 1972; Howie, 1976; Whalen and Xu, 1992; Liu and Samuel, 2004, and others), but this study's focus is in the presentation of everyday conversational utterances under as natural a circumstance as possible while maintaining reasonable control and consistency in the collection of the responses.

Mandarin was chosen for the language of study in order to assess the pitch contribution to comprehension in a tone language. In addition to native speakers being readily found, the functional load of tone in Mandarin phonology (the proportion of total phonological contrasts maintained by tonal categories) has been calculated and shown to be as great as that of vowels (Surendran and Levow, 2003), so there is no question of tone's importance in the formal organization of the language.

As an example of these tone contrasts, pronouncing [p^{h}a] with a quickly falling pitch through the syllable means ‘to fear’ and is written ⟨怕⟩, while the same syllable ([p^{h}a]) with a rising pitch means, coincidentally, ‘to climb’, written ⟨爬⟩. Without the tones, these two words would be homophonous in the language, so the contrast is maintained by tone, just as vowel categories maintain the contrast between [ti˥] (high level tone) ‘to kick’ and [ta˥] (high level tone) ‘him; her; it’.

Mandarin has a system of 4 “full” tones, and the ‘neutral’ tone which is a weak form of tone which occurs in unstressed syllables and coarticulates heavily with other tones (Cao, 1992; Chen and Xu, 2006). The tones have come to be identified by native speakers as simply “tones 1-4” (or sometimes the “high tone”, “rising tone”, “dipping tone”, and “falling tone” respectively), and “the neutral tone” or “clear tone”. The tones are predominantly distinguished from each other by the pitch contour produced over the voiced portion of a syllable (Chao, 1965; Duanmu, 2007; Lin, 2007).

Schematic pitch contours for each of the four full tones are shown in Figure 3.1. Tone 1 is a flat pitch contour, tone 2 is a rising pitch, tone 3 dips down and then rises back up, and tone 4 is a falling pitch. The neutral tone appears to have a pitch target in the middle of speakers’ pitch ranges (Chen and Xu, 2006), but its actual pitch contour varies greatly under the influence of neighboring tones.

Thus, Mandarin is a language which relies heavily on these tone categories which are defined in terms of pitch contour, and the present study measures how badly comprehension will suffer when pitch information is removed from the signal, without the other exaggerated compensatory acoustic cues that speakers might make when producing natural whisper (Liu and Samuel, 2004). LPC resynthesis allows us to remove the pitch information without changing other properties of the speech. This will tell us whether pitch really is vital in practice, or whether listeners can understand

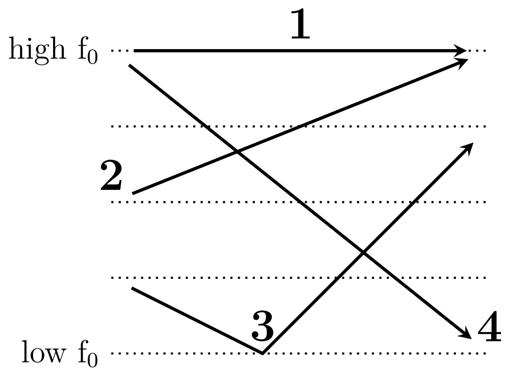


Figure 3.1: Schematic of the standard pitch patterns of tones 1–4 as conceptualized by native speakers, and taught to children and L2 learners. The neutral tone (not shown) meanwhile, is relatively short and varies widely in its pitch realization according to its neighboring tones (Cao, 1992; Chen and Xu, 2006). Horizontal dotted lines represent a 5-level division of a speaker's normal speaking pitch range. This visual scheme was the basis of the current IPA standard notation for tones, devised by Chao (1930). This figure is reproduced here for reference.

the speech just as well using other acoustic cues that covary with the pitch patterns of tones, such as duration (Blicher et al., 1990; Chang and Yao, 2007; Cabrera et al., 2014), voice quality (Yu, 2010; Bissiri et al., 2014), intensity contours (Whalen and Xu, 1992; Liu and Samuel, 2004), and segmental differences (Xu, 1997; Hu, 2004).

If the longer context of the full utterances and their semantic coherence (as well as distributional knowledge of the language such as the predictability of words, the statistics of the pairing of tones with syllable shapes, etc.), listeners may not suffer much with the removal of pitch information since the words occurring are highly constrained within the context of the utterance. If, on the other hand, tones are vital to the discrimination of otherwise homophonous competing words, and conversational reduction of segments makes segmental syllable specifications more ambiguous, listeners may struggle to keep up with the conversational utterances and their error rates (edit distances) will increase.

3.2 Methods

3.2.1 Overview

In order to measure the contribution of pitch to everyday Mandarin conversational comprehension in situ, in this experiment 30 native Mandarin listeners (most were the same subjects as in the perception experiment in Chapter 2) transcribed 50 conversational utterances prepared in such a way that some utterances provided pitch information to the listeners, and some did not (this “preparation” is described below in this same section). The difference between these two kinds of preparation will enable the direct comparison of listeners’ comprehension with and without pitch, and reveal how vital the pitch cues are themselves in Mandarin and how well listeners can do without them, using other acoustic cues present. In order to make this comparison, the stimuli need to be comparable except that in one case, pitch information is present, and in the other, pitch information is missing in the acoustics of the recording[3]. Linear Predictive Coding (O'Shaughnessy, 1988) resynthesis, described below, allows just that.

The method that will be used involves a set of conversational utterances, a resynthesis technique for removing the pitch information from the signal, gold standard transcriptions agreed upon by two native speaker colleagues, and the response transcriptions by the 30 participant listeners. The transcription task itself simply means that listeners hear a conversational utterance played through some headphones, and they type out the Chinese characters of the words that they heard in the utterance as accurately as they can. Then the degree to which these transcriptions are different from a gold standard transcription is an approximation of perceptual or comprehension error, i.e. if the gold standard transcription is ⟨最便宜的也都要十几 块⟩ but a listener types ⟨最便宜的也都要十七 块⟩, where the listener's second to last syllable ⟨七⟩ is transcribed differently from the gold standard ⟨几⟩, this indicates that they

[3] Individual utterances may vary in the predictability and therefore the ease of recognition of words, segments, and tones in their context, but utterances in this study are matched and counterbalanced across the two resynthesis preparations, so differences in predictability cannot explain differences between the two resynthesis contexts.

may have heard that word incorrectly, or, if the character they typed is homophonous with the gold standard character, they simply misunderstood the meaning, i.e. which morpheme (written character) among the homophones was intended. In addition to simply the Chinese characters that listeners directly transcribed, the data processing described below also enables the comparison of the segments and tones associated with the transcribed characters with the segments and tones associated with the gold standard transcriptions, revealing which consonants, vowels, and tones the listeners heard well and which were misheard as something else. In the example just given, the gold standard character is pronounced jǐ [dʒiː˨˩˦], while the hypothetical listener transcribed a character pronounced qī [tʃiː˥], so once the pronunciations are derived from the character strings, it becomes clear that the listener misperceived the voicing of the initial affricate and the tone of the syllable, but not the place or manner of articulation of the consonant or the identity of the vowel, reasonable mistakes to expect if pitch information is not present in the signal. Of primary interest will be how well the listeners were able to hear the segments and tones when pitch information is not audible in the recording, and how much worse this is than in the case of the full acoustic signal complete with pitch. To begin to prepare this method of comparison, a set of real conversational utterances were recorded and processed.

3.2.2 Materials

First, in order to create the stimuli for the transcription task, five Beijing-area native Mandarin speakers, 3 female, 2 male, were each recorded in conversation with a close friend in a noise-attenuating recording booth at either the Chinese Academy of Social Sciences, Beijing, or at the Douglass Phonetics Laboratory at the University of Arizona, Tucson. All speakers were college students, and grew up in the Beijing area or its neighboring Shandong Province until age 19 or later. One additional speaker was recorded for use in the practice stimuli as well. The recordings were made in the summer and fall of 2013 with a Countryman E6 omnidirectional head-mounted microphone positioned on the ear opposite the telephone. Recordings were not made through the phone; it only served to provide a context for the conversation. The conversations were completely self-guided. No topics were offered. Subjects were instructed to simply chat as usual. Only the participant on site was recorded, and their telephone interlocutors were informed of this at the beginning of each recording. In Beijing, the microphone was powered by a Shure FP23 preamplifier, and recorded on a Sony PCM-M10 solid-state recorder. In Tucson, the microphone feed went to a Symetrix 302 preamplifier, and then to an Alesis ML-9600 hard-disk recorder. Both recording setups were chosen for their flat frequency responses, low noise, and minimal distortion.

The purpose of this study is to capture real conversational speech comprehension to the extent possible in the laboratory. Care was taken not to artificially select stimuli of any particular difficulty or content, but to sample as representatively as possible from actual conversations. At the same time, this experiment is not a memory test, and transcription of words is a different task from merely understanding speech in a conversation, so overly long stretches of speech were avoided. From each of the five recordings, utterances (defined as turns at talk of 10 - 40 syllables) were selected at random using a script to automatically position a Praat (Boersma and Weenink, 2015) editor window at a random time point within the recording. The nearest utterance was selected, and the process repeated until 10 utterances had been chosen for the speaker. The resulting sample of 50 utterances (5 speakers, 10 utterances each) ranges from 10 - 35 syllables (mean 20, sd 4.8), from 2.3s - 5.0s in duration (mean 3.6s, sd 0.81), and from 3.2 - 8.4 syllables per second speech rate (mean 5.6, sd 1.3; figured from orthographic syllables of the gold standard transcriptions and the raw durations of the utterances, including pauses).

Each of the 50 utterance recordings were prepared to create two stimuli, a “full” stimulus and a “whisper” stimulus. Both stimuli were resynthesized in Praat (Boersma and Weenink, 2015) by Linear Predictive Coding (O’Shaughnessy, 1988), a mathematical model which functionally separates the acoustic contributions of the vocal source and the supraglottal vocal tract filter. Resynthesis was performed with a 5000 Hz spectral cap, 20 pole spectral matching, a 25ms analysis window with a 10ms step, and pre-emphasis of 6dB/octave beginning at 50Hz. The “full” stimuli are fully resynthesized versions of the sentences[4], while in the “whisper” resynthesis, all vocal pulses are removed from the source to simulate whispered speech, but without the potential compensatory effects of natural whisper (Liu and Samuel, 2004; Chang and Yao, 2007; Li and Guo, 2012). Following resynthesis, the original intensity curve was reapplied so that both the “full” and “whisper” versions had the same intensity features. This process generated 100 stimuli, 10 “full” and 10 “whisper” from each of 5 speakers. Sample “full” and “whisper” versions of an utterance are shown in Figure

[4] This neutral resynthesis is intended to retain comparability of all presented stimuli in terms of their processing.

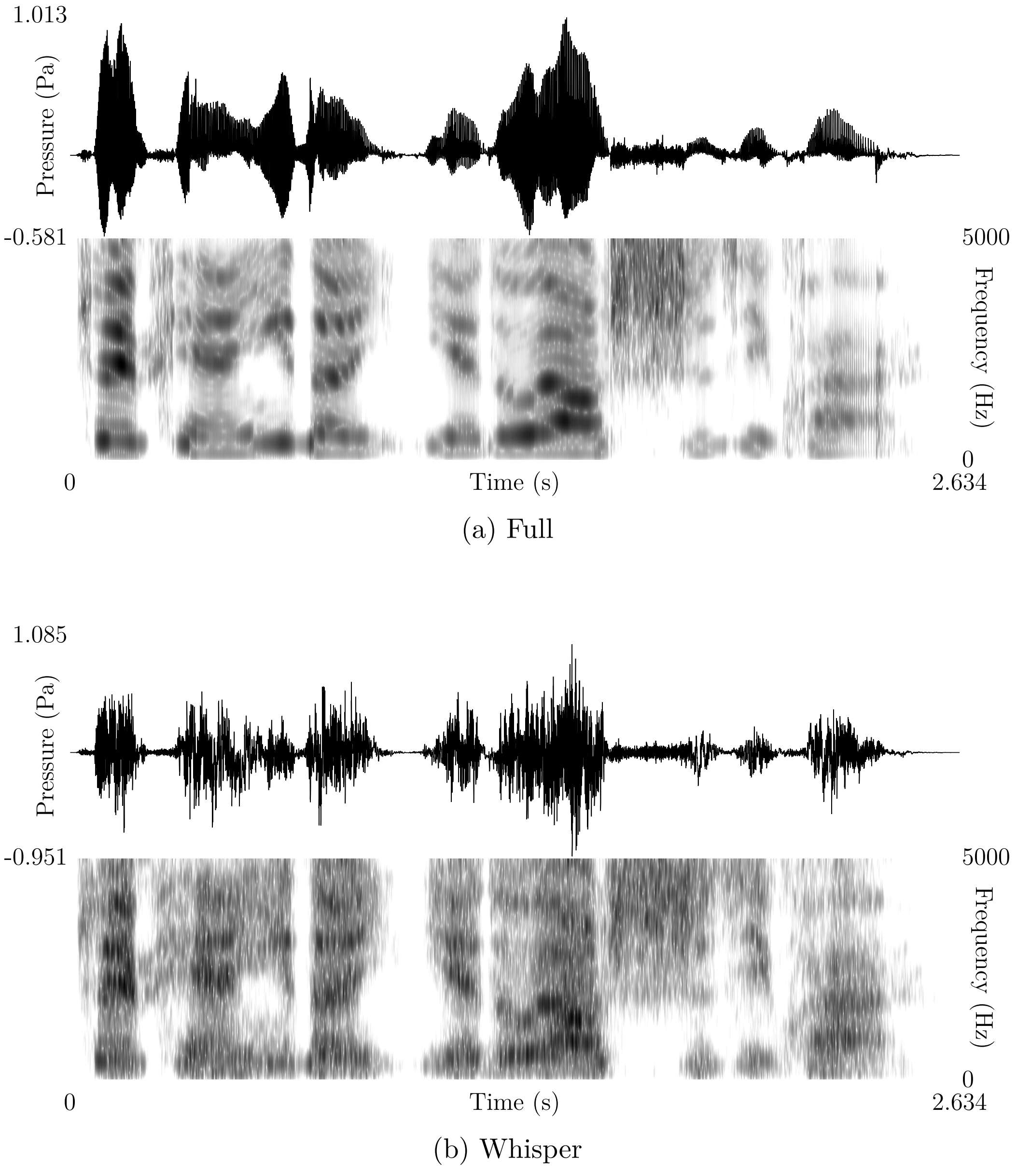
3.2.

3.2.3 Subjects

The listeners were 34 native Mandarin speaking students at the University of Arizona (20 female, 14 male; ages 18-29, mean age 23, st.dev 3.4). They participated in Summer and Fall of 2014. All reported normal hearing, reading, and language function, and all had grown up in mainland China until at least the age of 16, but their hometowns within China varied greatly. They had spent between 3 months – 7 years in the U.S. at the time of testing (mean time in U.S. 2.1yrs, st.dev 1.7). The participants were compensated monetarily for the roughly one hour that they spent in the lab transcribing the conversational utterances.

3.2.4 Procedure

Stimuli were presented to listeners using Praat in two counterbalanced lists. Instructions were entirely in Mandarin, with text within the experiment prepared with the help of a native Mandarin-speaking colleague. In each trial, the resynthesized utter-



(a) Full

(b) Whisper

Figure 3.2: ⟨最便宜的也都要十几块⟩ zùi piányi de yě doū yào shíjǐ kuài ‘Even the cheapest ones cost more than ten dollars.’

ance recording was played twice, after which the subject was given a text box to type out what they thought had been said, utilizing a standard Mandarin pinyin input method editor (“SunPinyin” in Linux, “Microsoft Pinyin IME” in Windows). The resulting text is a string of Chinese characters. After submitting their transcription, they were played the utterance a third time, and then were shown what they had typed and were given an opportunity to correct or add to their transcription before moving on to the next item. This final, corrected Chinese character string constitutes the response. Instructions were to type out exactly what the person said, taking care not to make typing errors. No time limit was imposed, but subjects were implicitly limited by their memory of the utterance and the three times they were exposed to the stimulus. Subjects heard 5 “full” utterances from each of the 5 recorded speakers, followed by 5 “whisper” utterances from each speaker, and the items were counterbalanced in two lists so that the items presented as “full” to half the listeners were presented as “whispered” to the other half. The full experiment presented 2 practice items (1 “full”, 1 “whisper”), 25 “full” items, and then 25 “whisper” items. In all, it took roughly 1 hour for subjects to complete. Despite the length, subjects generally reported enjoying listening to this casual conversation, and many clearly showed high confidence that they had performed the task well.

3.3 Results

3.3.1 Data preparation

The analysis for this experiment is designed to enable the comparison of subject responses to gold standard transcriptions agreed upon by two native colleagues under ideal listening conditions. The intended comparisons are at the Chinese character level (which subjects directly entered), the segment level, and the tone level. The dependent variables (DVs) for analyses in this experiment are Levenshtein edit distances (Levenshtein, 1966; Day, 1984) of responses from their corresponding gold standard transcriptions. These text-based differences are not indicative of "perceptual distance" of units, but serve to tally the numbers of errors listeners make, however perceptually distant the mistakes are. The units of comparison are characters, tones, and segments. In order to associate the intended segments and tones to the transcribed Chinese characters, a dictionary was prepared to reference tones and segments in. And because standard Chinese text does not contain spaces or otherwise indicate word boundaries, parses were made of each Chinese character response given by subjects. The "responses" are the final transcriptions provided after the third time an utterance was played, and the subject had an opportunity to revise their original entry.

The dictionary employed for this analysis is the LDC Mandarin Lexicon (Huang et al., 1997), created to cover the training set of the Mandarin CallHome corpus (Canavan and Zipperlen, 1996) using a rule-based parsing algorithm. It contains 44,405 headwords, and provides 98% coverage of the CallHome `devtest` recordings, an independent sample of 20 telephone conversations. The dictionary provides Chinese character headwords, pinyin orthographic representations (Murthy et al., 1979) including tone markings, tone sequence of the headword syllables with tone sandhi applied (Chen, 2000), phonetic transcriptions, the frequency of the word "in 3,431,707 words of Xinhua newswire"[5] (Huang et al., 1997), and the frequency of the word in the CallHome training set (Canavan and Zipperlen, 1996, 155,276 tokens, 80 telephone conversations). The dictionary headwords are used for parsing subject response transcriptions, and its tone and segmental information is used here to compute distances

[5] No further detail is given regarding the source of these frequencies, but they are not relevant to the current study.

of listeners' transcriptions from a gold standard transcription.

A native Mandarin speaking colleague kindly provided Chinese character and also pinyin transcriptions (indicating segments and tones, and also word boundaries) of each of the stimulus utterances with the full recordings (the entire conversations) available to her for context. A second Chinese colleague confirmed the transcriptions of the first. Segment transcripts were then created using the pinyin-to-segment correspondences described in the documentation of the dictionary. These character, pinyin, and segmental transcripts for each utterance serve as the gold standard transcriptions for this analysis.

In the handful of words in the gold standard pinyin transcriptions provided by my native speaker informants differing from the dictionary pinyin for those words, the dictionary version was retained for consistency since the dictionary entries were the basis for the response parsing. E.g. the dictionary pinyin péngyou 'friend' was given by the native speakers as péngyoŭ; the dictionary pinyin wănshang 'evening' was given as wănshàng; etc.

Some words in the gold standard transcriptions and in listener responses were not present in the dictionary, even when decomposed completely into single characters.

These were added to the lexicon by referencing another dictionary, the CASS (2012), in order to insure that all responses and gold standard transcriptions are represented in the dictionary headwords. These supplementary words are listed in Appendix A.

Even with the LDC Mandarin Lexicon and the fact that it is based on the same speech style and context as our stimuli, even with the excellent coverage of dictionary headwords over novel conversations, and even with the appropriate kinds of tone and segmental information provided, there are significant challenges to attributing a certain parse, a certain tone sequence, or a definite segmental string to a Mandarin free response. For example, there are 1,024 unique ways of grouping the characters of the item ⟨就是一个寓言故事就是北风和太阳的故事你应该知道吧⟩ into words indexed in the dictionary (in this case, simply meaning groupings of characters into constituents which occur as headwords). For example, ⟨就⟩ is a headword, as is ⟨是⟩, as is ⟨就是⟩. Once a parse is identified, there is further variation in pronunciations of a given word, e.g. ⟨那⟩ can be pronounced nà [nɑ˥˩] ‘that’ or nèi [neɪ˥˩] ‘that’; ⟨边⟩ can be pronounced biān [pjɛn˥] or bian [pjɛnɸ]; ⟨着⟩ can be pronounced zhāo [tʃau˥] ‘a move, a trick’, zháo [tʃau˧˥] ‘to touch’, or zhe [tʃɤɸ] (progressive aspect particle). These many possibilities can make a Mandarin string of characters highly ambiguous.

For each response, a complete parse tree was constructed by a recursive algorithm using dictionary headwords. The resulting tree representation contains all possible unique ways of analyzing the response completely in headwords. For example using the item mentioned just above, one branch of the tree would begin with the headword ⟨就⟩ and one would begin with ⟨就是⟩, and each of these branch again each time multiple headwords match the beginning of the remainder of the response. The resulting parse trees contain between 1 and 8,192 parses at the character level.

To regularize the texts in preparation for parsing, punctuation, special characters, and the dialectally variable "erhua" suffix ⟨儿⟩ er [ɹ] (in words such as ⟨里边儿⟩ lǐbianr [lipiaɹ̃] 'inside'; Chao (1965); Duanmu (2007); Lin (2007)) were removed from both the character and pinyin transcriptions in the responses and gold standard parses prior to deriving the segmental transcripts. In practice, these were removed simply by deleting word-final ⟨儿⟩ characters, which would have the unintended side-effect of changing words like ⟨女儿⟩ nǚěr[6] 'daughter' to simply ⟨女⟩ nǚ 'female', but there were no occurrences of words like this in the stimulus utterances.

In preparation for the segment and tone comparisons, the character parse trees

[6] The pinyin letter "ü" indicates the rounded high front vowel [y].

were then expanded in a similar process utilizing the pinyin entries in the dictionary for each of the words in each parse, so that one pinyin parse might have nà, while one might have nèi, because those are both possible pronunciations of the character ⟨那⟩. From these pinyin trees, the tones were extracted for the tone comparisons, and the pinyin segmental characters were translated to an ASCII segment transcription (roughly corresponding to IPA transcription) for the segment comparisons, using the segment correspondences in the dictionary. In other words, pinyin syllables in the dictionary all also have segmental transcriptions.

With the comparison elements compiled, a Levenshtein edit distance (ED) was computed from each response parse to the gold standard, and the minimum distance of all was taken as the distance for the response. The Levenshtein Distance (Levenshtein, 1966; Day, 1984) is a measure of difference between strings, the minimum number of unit edits (adding a unit, removing a unit, or substituting one unit for another) necessary to transform one string into another. For example, the string "those" is 2 character edits' distance from "this", changing "i" for "o", and adding "e" to the end. Or as a simple Mandarin example, if the gold standard utterance were ⟨有些店就没有⟩ yǒuxiē diàn jiù méiyoǔ 'Some stores just don't have it' (tones: 31 4 4 23;

segments: `yowxyE dyEn jyow meyyow`), and the response were ⟨有鞋子店？没有⟩ yǒu xiézi diàn? méiyǒu ‘Is there a shoe store? No, there aren’t any’ (tones: 3 20 4 23; segments: `yow xyEzI dyEn meyyow`), the character level ED is 3 because the response exchanges ⟨些⟩ for ⟨鞋⟩, adds the character ⟨子⟩, and omits the character ⟨就⟩; the tone level distance is also 3, because the response tones replace the second tone, tone 1, with tone 2, further add a tone 0 following that, and omits one of the consecutive gold standard tone 4s; the segment level ED is 6, because the segment transcription contains the two added segments `zI` between `xyE` and `dyEn`, and further omits the four segments `jyow` between `djEn` and `meyyow`. The ED in all cases is minimally the difference in length between the two strings (if the response contains all correct units, but either omits or adds some segments relative to the gold standard), and maximally the length of the longer string (if the response contains no correct segments at all). The ED is computed here with a standard dynamic programming algorithm illustrated in Table 3.1. This approach can be used with different constituents, e.g. words, letters, syllables, tones, features, etc. In the present case, the edit distance was computed independently for each comparison unit (for the character string of the response, the tone sequence parses derived from the pinyin parses for the response, and the phonetic segment parses derived from the pinyin parses) so that the particular parse leading to a minimum distance for tone need not be the same parse which leads to the minimum distance for segments, but is considered on the basis of the unit at issue. Finally, the edit distances were normalized for the item length, dividing by the number of units (characters, tones, segments) in the gold standard, yielding an "edit distance-per-unit" average.

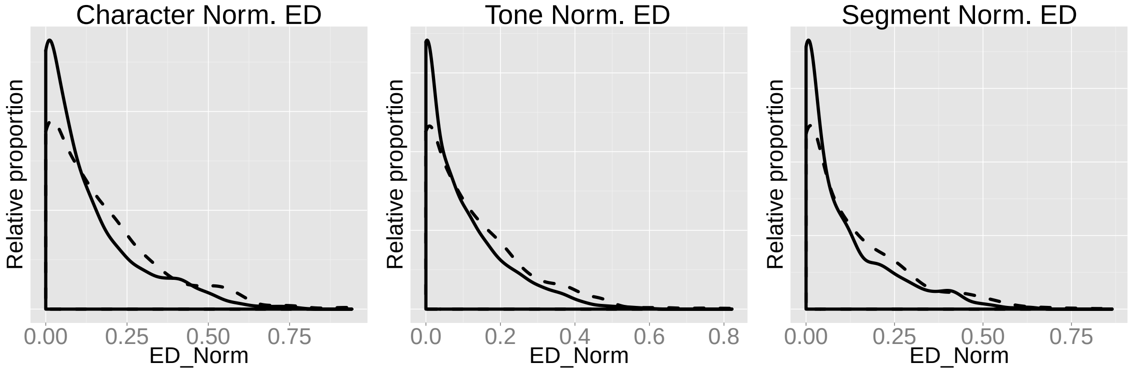
3.3.2 Summary Statistics

The distributions of normalized edit distances (average edits per unit, similar to proportion error) for full and whispered utterances in the three units (Chinese characters, tones, and segments) are shown in Figure 3.3. The Chinese character distances approximate the proportion of morphemes misheard/misinterpreted, the tone distances approximate the proportion of tones misidentified, and the segment distances approximate the proportion of segments misperceived. In each case, the individual distributions are not of paramount importance, since no listeners are perfect, and we expect a variety of error rates with different items and listeners, but rather the feature of interest is the *difference* between the full and whispered item distributions,

		Response																
		∅	北	京	酒	吧	又	分	闹	吧	和	又	分	京	吧	什	么	的
Gold standard	∅	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	北	1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	京	2	1	0	1	2	3	4	5	6	7	8	9	9	10	11	12	13
	那	3	2	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	酒	4	3	2	1	2	3	4	5	6	7	8	9	10	11	12	13	14
	吧	5	4	3	2	1	2	3	4	4	5	6	7	8	8	9	10	11
	又	6	5	4	3	2	1	2	3	4	5	5	6	7	8	9	10	11
	分	7	6	5	4	3	2	1	2	3	4	5	5	6	7	8	9	10
	闹	8	7	6	5	4	3	2	1	2	3	4	5	6	7	8	9	10
	吧	9	8	7	6	4	4	3	2	1	2	3	4	5	5	6	7	8
	又	10	9	8	7	5	4	4	3	2	2	2	3	4	5	6	7	8
	分	11	10	9	8	6	5	4	4	3	3	3	2	3	4	5	6	7
	静	12	11	10	9	7	6	5	5	4	4	4	3	3	4	5	6	7
	吧	13	12	11	10	7	7	6	6	4	5	5	4	4	3	4	5	6
	什	14	13	12	11	8	8	7	7	5	5	6	5	5	4	3	4	5
	么	15	14	13	12	9	9	8	8	6	6	6	6	6	5	4	3	4
	的	16	15	14	13	10	10	9	9	7	7	7	7	7	6	5	4	**3**

Table 3.1: The Levenshtein distance is computed using a tabular data structure, in this case comparing a Chinese character response to its corresponding gold standard, ⟨北京那酒吧又分闹吧又分静吧什么的⟩ běijīng nà jiǔbā yòu fēn nàobā yòu fēn jìngbā shénmede ‘Beijing’s bars are divided into “loud bars” and “quiet bars”, or something like that.’ The table is filled from top left to bottom right. At each cell we enter the smallest of its top, leftward, or top-left diagonal squares if the symbol of the row is equal to the symbol of the column, or 1 more than that number if the two symbols are not equal. Moving rightward in the table corresponds to an added symbol in the response, moving downward corresponds to the omission of a symbol in the response, and moving diagonally down and to the right corresponds to a matching symbol (with no penalty) or a substituted symbol (we add 1). The number in the bottom righthand corner becomes the final distance, the optimal minimum number of edits between the strings. The gray path indicates a minimum distance path (there can be more than one).

since this is the indicator of how much listeners are relying on pitch for that category of unit.



(a) Characters (b) Tones (c) Segments

Figure 3.3: Density plots of normalized Levenshtein edit distances of responses to "full" (solid line) and "whispered" (dashed line) utterances **a)** by Chinese character; **b)** by tone; and **c)** by segment. The x-axis represents average edit distance per unit; the y-axis presents the proportion of total items at each x value. The large peak on the left side indicates that a large proportion of the items had very low average EDs. The area under the curve represents all items in that resynthesis condition, and the height at each x value conveys how many of the items had EDs near that value.

In all three distributions, the whispered utterance distributions are flatter and spread out more to the right, meaning the transcriptions in whispered utterances were further from the gold standards, i.e. had higher error rates on average. Subjects recovered tone categories and segmental categories well, despite the loss of pitch. The character distribution is somewhat flatter (less sharply pointed on the bottom ED end) than the other analysis units in both resynthesis conditions, since it is a larger unit and subsumes the two other types of errors (errors of tone or segment will very likely lead to a character error, on top of which even homophones may be typed with the wrong characters even without any tone or segmental errors). Nevertheless, all units have strong peaks at the low-ED end, which indicates low error rates in all cases.

3.3.3 Linear Mixed-Effects Modeling

Model Selection This experimental design involved three random effects: Subject (i.e. listener; 1–34), Speaker (1–5), and Item (i.e. utterance; these are nested within speakers). For this reason Linear Mixed-Effects (LME) models were used for their flexibility in model specification and control for the random effect structure[7], and were used to model the effect of resynthesis condition on the error rates of responses with respect to characters, tones, and segments. The model fitting procedure was the same in each of these three cases. It began with a maximally specified model with dependent variable Normalized Edit Distance, the fixed effect Resynthesis condition (with "full" as the reference level), and random effects Subject, Speaker, and Item, with Items

[7] Once again, in Chapter 2 ANOVA was used instead because of model fitting time and convergence problems. That experiment had a simpler random effects structure however, which makes ANOVA an adequate choice.

nested within speakers, and with random intercepts and slopes for all random effects. This is the initial maximal model. Barr et al. (2013) show through Monte Carlo simulation that LME models generalize best when the maximal random effects structure justified by the experimental design is used in the model. In this case for our three models, the maximal structure includes intercepts and slopes by resynthesis condition for the random effects Subject and Items within Speakers. Speakers themselves also have individual random intercepts. Log-likelihood nested model comparisons were made checking each individual slope and intercept by removing them each in turn, and checking that the model significantly better fits with the term retained, and in all cases but one, the inclusive model with the maximal random effects structure was significantly better fitting than models with any of the random effects slopes or intercept terms omitted. The one exception was in the character model. For this measure, the inclusive model with random slopes for resynthesis condition by Subject was not significantly better fitting than the model with random intercepts only (χ^2=4.61, χ^2 df=2, p<0.1), but here we will follow Barr et al. and retain the random slopes as well. The pattern of results is the same with or without, and effect magnitudes are very similar also.

Having arrived at the model structure, one last check was performed. The dependent variable Normalized ED implicitly contains an effect of utterance length, since it has been normalized by dividing by the number of units in the utterance, but it is possible that the effect of utterance length is different in the two resynthesis conditions, “full” and “whisper”, since the amount of acoustic information provided may interact with the encoding of the signal in perception, affecting memory and the error rates of responses. In order to verify that there is no interaction of the effect of length with resynthesis condition, the same models were refit with raw edit distances as the dependent variable, and Resynthesis condition and utterance Length as fixed effects, and a nested model comparison confirmed that a model including an interaction term did not significantly improve the model fit. This was true for all three unit models, the character model, the tone model, and the segment model.

Model Results Summary distributions of normalized edit distances are plotted above in §3.3.2. Summary by-category bar plots are shown in Figure 3.4. The character unit model showed a significant nonzero intercept for the fixed effect Resynthesis condition (β=0.124; standard error=0.021; df=7.2; t=6.01; p<0.001), demonstrating that we have sufficient power to detect the error rates. Subjects are not at ceiling performance, even in the “full” reference level. The “whisper” effect also showed significantly higher normalized error rates (β=0.036; standard error=0.008; df=41.6; t=4.33; p<0.001).

The tone and segment unit models echoed the pattern for characters. The tone model also revealed a significant nonzero intercept representing the “full” condition normalized edit distance (β=0.089; standard error=0.015; df=6.5; t=5.96; p<0.001), as well as a significantly higher error rate for “whisper” (β=0.031; standard error=0.007; df=46.4; t=4.43; p<0.001). Likewise the segment model “full condition” intercept (β=0.098; standard error=0.018; df=6.0; t=5.54; p<0.01) and the “whisper” level effect (β=0.028; standard error=0.007; df=42.3; t=3.92; p<0.001) were significant.

Figure 3.4 shows the effect sizes and 95% prediction intervals for the effect of resynthesis in each of the three unit analyses. In each case, the error increase due to the loss of pitch information is between 29-35%.

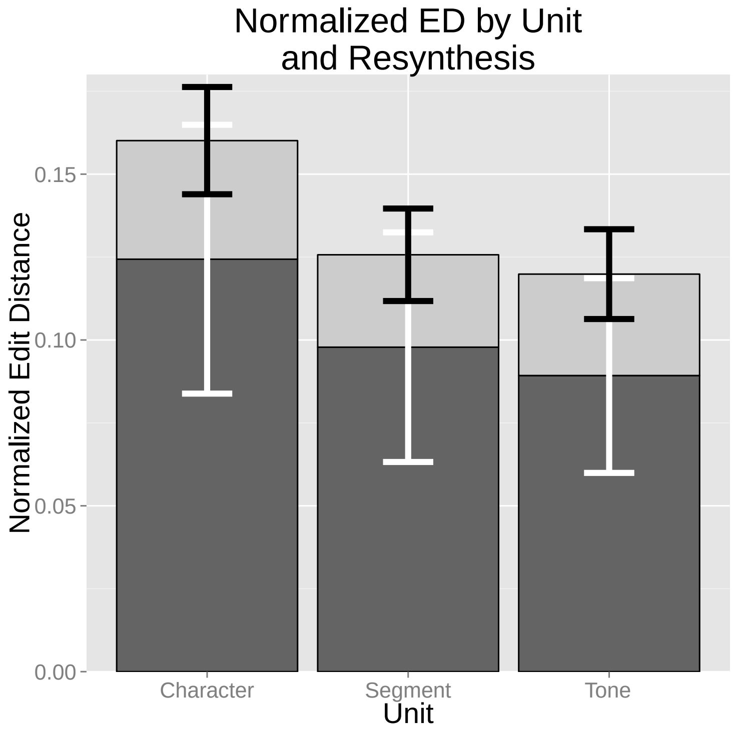


Figure 3.4: Normalized edit distance (edits per unit length) for characters, tones, and segments. Lower dark bars indicate error rates in the “full” resynthesis condition. Light gray bars above them show the increase in error rates in the “whisper” condition. Error bars indicate 95% prediction intervals for each.

3.4 Discussion

The results of this study have a clear message as regards the importance of pitch information in realistic conversation for Mandarin listeners. A loss of pitch information causes a one-third increase in listening errors both in terms of tone categories, segments, and by a more direct measure of overall semantic comprehension, characters. Since Mandarin writing is a nearly morphographic system in which characters directly represent morphemes, the character level results can be considered an approximation of the general semantic comprehension of the listeners. A botched character is likely to indicate that the content of the utterance was not fully received, even if the character written had the same pronunciation as the correct response.

For example, ⟨你还在你那个房子里住着? 你室有， 还是 那一个人吗?⟩ nǐ hái zài nǐ nàge fángzi lǐ zhùzhe? nǐ shì yǒu, háishì nà yīgè rén ma? ‘You are still living in that house? Does your room have (one), or is it that one person?’ was one response to the item ⟨你还在你那个房子里住着，你室友还是 那一个人吗?⟩ nǐ hái zài nǐ nàge fángzi lǐ zhùzhe? nǐ shìyoǔ hái shì nà yīgè rén ma? ‘You’re still living in that house? Your roommate, is it still that one person?’. The differences between the two are underlined. In the character transcriptions, the listener wrote ⟨室有⟩ rather than ⟨室友⟩. Even though these two strings are pronounced identically, shìyǒu, with the same tones and the same segments, they differ in their word parsing, since the listener's ⟨室有⟩ is two words, ⟨室⟩ shì 'room', and ⟨有⟩ yoǔ 'have', while the original is a single word, ⟨室友⟩ shìyoǔ 'roommate'. In this case, the character level distance is 1, since ⟨有⟩ is not ⟨友⟩, and this reflects that the listener misconstrued the word 'roommate'. The mistaken interpretation is also facilitated by another homophonous pair, the listener's ⟨还是⟩ háishì 'or', versus the original ⟨还⟩ hái 'still' and ⟨是⟩ shì 'is'. In this latter case, because the two are also written identically, there is no additional edit distance resulting. As for the tone and segment edit distances, since the listener's response has the same tones and segments as the original, the tone and segment edit distances are zero, i.e. all the tones and segments appear to have been correctly perceived, but the error made was a misinterpretation at the word parsing level.

It is important to note that the removal of *pitch* from the speech signal is not the same as removing *tone* from the signal. Tonal categories are manifested in other aspects of the acoustic signal such as duration (Blicher et al., 1990; Chang and Yao, 2007; Cabrera et al., 2014), intensity (Whalen and Xu, 1992; Liu and Samuel, 2004), voice quality (Yu, 2010; Bissiri et al., 2014), and vowel quality (Xu, 1997; Hu, 2004), which perhaps partly explains the reasonably good performance of listeners in both resynthesis conditions. Nevertheless, removing pitch information resulted in a large increase in error rates, which means that listeners struggle to recover the information in the high variability of the available cues in conversation.

Because the effect of pitch elimination is consistent across the three units of comparison, it appears to cause a general issue for perception and comprehension rather than one targeted at tone categories. Cutler and Chen (1997) found that tone perception is slower and more error prone than segmental perception, but in these results, in practice, it does not appear that tone errors are more likely than segmental errors. This could be an artifact of perplexity differences, since there are fewer tone categories to choose from than segment categories, and therefore chance level tone guessing is more likely to be correct than equivalent segmental guesses. Therefore, even though tones appear to have similar error rates as segments, this is simply a case of the balancing of fewer categories to select from and the more error prone perception in the case of tones.

The “full” condition results expose a limitation of the task as it was presented. It is difficult to believe that Mandarin listeners normally misperceive one in twelve or so morphemes (or segments or tones) in practice. In some senses, the task here was more difficult than in normal conversation, as there were only isolated utterances out of their conversational context, and with this comes a certain extra challenge as listeners intently seek to orient themselves within the conversation from scratch with each utterance, as if suddenly thrown into a conversation in the middle. This situation is certainly not unusual to encounter, however, as when joining a gathering or meeting late, turning on the television or radio in the middle of a program, etc. So while the disjointed or sudden nature of conversations isn’t at all unnatural, it is somewhat intensified since each utterance is quite independent of all the others.

Another likely contributor is simply typographical error. The common pinyin input method editor predictively offers candidate Chinese characters for the pinyin the user enters, and there is some evidence in the responses that listeners may have sometimes chosen the default first choice in the list of offered candidates in places where they meant to select another option. For example, if beginning a sentence one types `xiaowei` intending the characters ⟨小薇⟩ ‘Xiao Wei (a name)’, the Linux Sun-

Pinyin IME offers the options ⟨校尉⟩ ‘military officer’, ⟨小伟⟩ ‘Xiao Wei (a different name)’, etc., in a list before the intended ⟨小薇⟩. In addition, the IMEs typically log the words used, so that if one selects a given word for some pinyin letters entered, this word is moved up the cue the next time those same pinyin letters are entered. So when users type, they can select words at the front of the selection cue simply because they assume it is the most likely word to appear at the front, or because the cue ordering has changed since they last entered the word, or simply from inattention to the choice. This means that to some extent, the dependent measures employed here are inflated by typing errors. Chen and Lee (2000) report an approximately 4.6% character error rate in their input training corpus for a pinyin input engine, for example. If this is roughly representative of the typing error contribution to the errors in this study also, the effect of losing pitch information is then proportionally even larger, around one half. To avoid this typing confound, however, one would have to custom code an input method editor which is not predictive, in order to force deliberate selection of each character, or to log the original pinyin the subject typed to access the characters, or else brave the time-consuming and error prone processing of hand-written responses.

Although in some ways the task in this study was more difficult than in an ordinary conversation, at the same time, there are ways in which it was markedly easier than under normal conversational circumstances. Listeners heard each utterance three full times through, in precisely the same way. And because of the three repetitions, they also had considerably more time than a usual listener would have in the stream of a conversation. And all of this occurs in a completely passive listening environment in which they do not need to plan any responses (apart from reproducing what they are listening to), signal anything to the speaker, navigate the structure or manage the trajectory of the interaction, or retrieve much topical information from memory. The project as a whole seems, then, on the balance, to be at least somewhat easier than usual conversation.

Although the functional load of tone in Mandarin phonology (Surendran and Niyogi, 2003) was computed some time ago and is powerful evidence of its vital psychological importance in the language, the present study provides a similar measure, and similar story, in the comprehension of everyday conversational utterances in a context much closer to actual usage. And although it reinforces that existing understanding, it expands the generality of its claims into a domain that people care about greatly, understanding one another. Students of Mandarin, computational linguists, and human language technology engineers should know that failing to accurately produce pitch contours —independently of the other concurrent acoustic cues to tone categories— will cause considerable extra comprehension errors for listeners even above and beyond what they can compensate for with extra effort, or, to some extent, context. Together with the other channels of acoustic information in speech, pitch provides a very great contribution to tonal, segmental, and overall comprehension.

Chapter 4

General Discussion

4.1 Summary of Results

This dissertation investigated how speakers of Mandarin produce and perceive tones in casual spontaneous speech. Previous work (Berry, 2009; Brenner, 2013) detailed the properties of the tones speakers produce in conversation, and how the cues to the tones diverge from what is normally expected in careful speech. The present Chapter 2 investigated how well native listeners are able to identify what tone they are hearing based on the acoustic cues present in spontaneous conversational vs. carefully pronounced speech. Chapter 3 investigated how pitch information in spontaneous speech contributes to listeners' understanding of entire utterances.

4.1.1 Tone Identification

In the perception experiment, listeners identified the tones of “full”, “hummed”, and “whispered” words. The reduced acoustic cues to tone uncovered in the conversational words by the acoustic study made tone identification more difficult for listeners, and they responded by combining the various available sources of information about tone in a complex way, using idiosyncratic combinations of cues depending on the tones.

In careful speech items, listeners were roughly equally proficient in identifying tone categories regardless of whether they had full acoustic information, only the hummed pitch contour for the syllable, or a whispered version of the syllable, indicating that the pitch contour itself (“hum”), and also the segmental level acoustic detail itself (“whisper”), are both sufficiently informative of tone categories for good tone identification.

In conversation, the reduced informativeness of individual cues makes it necessary to integrate multiple sources of acoustic evidence, and creates a greater hardship when sources of that evidence are limited or withheld since the cues available for the “triangulation” of tones become scarce. This is similar to the diagnosis of illness. If the symptoms are acute and closely associated with the illness presented, for example a heart attack, just one or two key symptoms may suffice to arrive at a diagnosis, and the diagnosis can be swift and reliable. If however, an illness is associated with many symptoms commonly presenting in other illnesses also, diagnosis will depend on the specific combination of many symptoms, and if any vital sources of information are not available, for example if the patient is unconscious and cannot answer questions, efforts to identify the cause may be much more difficult and error prone.

At the same time, just as in medical diagnosis, if one is forced to combine many sources of cues, the resulting “diagnosis” may sometimes reveal more about the underlying cause than if one or two cues seem by themselves to settle the matter. For example, listeners in this study were more able to distinguish tone 3 syllables from tone 2 syllables prior to another tone 3 (i.e. in the Sandhi environment) in conversational items than they were in the careful speech tokens. At least in some cases, diminished primary cues that signal a tone 2 allowed listeners to make a more diversely informed characterization of the tone, revealing that it was indeed an underlying tone 3.

4.1.2 Dictation

The dictation experiment presented another case where many cues were available to listeners in concert. In both resynthesis conditions, when listeners heard the full signal, or when they heard only a whispered version of the utterance, they performed fairly well in their transcription accuracy in terms of the tone sequences of their responses (8.9% in the "full" resynthesis condition, 12.0% in "whisper"), the segments in their responses (9.8% "full", 12.6% "whisper"), and the overall comprehension of the utterances as indicated by the characters they entered (12.4% "full", 16.0% "whisper").

Even though in the "whisper" context the pitch contour had been excluded from the resynthesis, the cues that remained in the utterance recordings in the syntax, duration, and intensity patterns, appear to be sufficient to enable reliable perception overall. At the same time, the pitch information is not simply redundant. There is a significant one-third additional error increase over the full resynthesis condition (again by all three units compared). Put another way, roughly 25% of tone category information is unrecoverable from other sources in the signal.

4.2 Comparison of Identification and Dictation Results

In the perception experiment, listeners heard isolated word list and conversational two-syllable words in “full”, “hum”, and “whisper” resynthesis styles, presenting full, pitch only, or pitch-absent information streams, to which listeners responded with the tone category (1–4) of the first syllable manipulated tone. In the dictation experiment, listeners heard entire conversational utterances in “full” or “whisper” resynthesis styles, presenting full or pitch-absent information streams, and transcribed the text of the utterance . From the response text, we inferred semantic comprehension, perceived segments, and perceived tones.

The perception experiment is aimed at revealing how listeners perceive the tone categories of words in isolation, using strictly the acoustic information provided in the three resynthesis styles, and comparing careful and conversational speech styles. The dictation experiment is rather designed to measure the contribution of pitch information within natural conversational utterances.

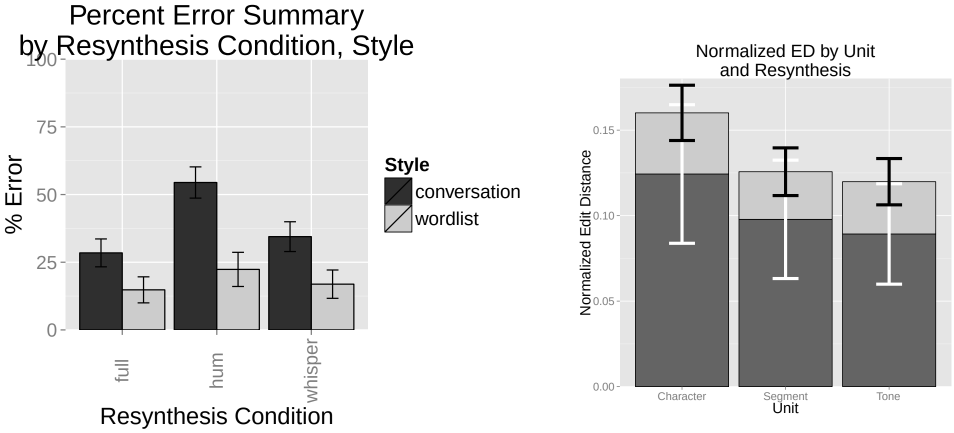


Figure 4.1: General summary of error results from the Perception experiment and the Dictation experiment.

4.3 Processing of Cues in Spontaneous Speech

Although the tasks involved in the two experiments were very different, the patterns across them are suggestive. In the perception experiment where subjects responded to isolated words, error rates are much higher in both the full and whisper conditions than the tone error rates in the larger scale context of the dictation experiment. Of the 29% or so tone errors made in the perception experiment with isolated words, in the dictation experiment with the longer context of full utterances listeners were able to recover roughly two-thirds of the correct tone categories.

Interestingly, the proportion of errors caused by the whisper condition is equivalent across the two experiments, increasing error rates by roughly one third relative to the full condition in both studies. This suggests that the information about tone present in the pitch contour is local to the words themselves, and not distributed far beyond their borders. Despite the fact that greater context lends better information about words and their tones, the information present in the pitch itself is proportionally equal at larger scales.

The increased multidimensional attention in conversation would seem to suggest it requires more processing and should take longer than careful speech in which a smaller number of cues more clearly signals speech categories generally. This expectation is borne out in studies like Ernestus et al. (2002), where reduced word forms were identified more slowly than those with stronger identifying cues.

In the perception experiment, the distributions of each of the cues had to be estimated under very challenging circumstances, from essentially random isolated words from word lists and conversation. As we saw in the acoustics study, the distributions of cues in conversation are far less informative of the tone categories, and yet even in the worst cases, listeners performed well above chance. It may be that the combinations of cues are more informative to the categories than is evident from the cues we measured. It is also likely there may be additional cues not anticipated in the existing literature, but the current project is not designed to expose those additional cues.

4.4 Implications for Second-language teaching of Mandarin

It may appear that the results here are indicative of a relatively minor role of tone in Mandarin word perception and general conversational comprehension. One may make one's own decision about whether a one-third increase in errors when pitch information is unavailable is important or not, but L2 learners and teachers of Mandarin should be reminded once again that the removal of pitch undertaken in these experiments differs from the removal of *tone*. Although the pitch information is absent, the other concurrent identifying cues to tone categories remain in the signal, so that the "whisper" presented is the whisper of a speaker who has produced the speech with tone categories in mind. Information identifying those categories must be considered to be largely intact in the associated cues. A speaker naïve to the proper production of the tones will likely not produce these tone category cues, but rather will produce contradictory or otherwise misleading and distracting cues not associated with the tone categories. In normal voiced speech this will be compounded by the tone-unrelated pitch contours. In this sense, the dictation experiment presents a kind of measure of the importance of pitch information in native speech rather than total tone information in Mandarin in general.

4.5 Future Directions

The research detailed here suggests a number of compelling questions for further exploration. The next steps will confirm results and expand the generalizability of findings from the present study.

The perception experiment presented stimuli from only a single speaker, and naturally there are questions of how generalizable the findings are to other speakers. There is bound to be some variation in the robustness of cues produced by different speakers, and perhaps in the way speech style (careful, conversational) is realized by different individuals. An exploration of which cues tend to vary by speaker and which are produced more consistently across speakers is likely to lead to a better understanding of which sources of information are more useful to listeners and which constitute perceptual noise. For example, if some speakers utilize a greater pitch range than others, it may be that pitch information will be of greater use to listeners, while speakers with a narrow range may induce in listeners attention to other streams of information such as segmental changes.

The task itself involved in the perception experiment also warrants further exploration. The present experiment utilized disyllabic words in order to offer some context to listeners for use in the slightly unnatural task of categorizing the tones of manipulated syllables. A comparison of two-syllable words, as in the present study, with a similar task using monosyllabic words would provide some insight into the role of the second syllable here. Using single syllable stimuli would force exclusive attention to the manipulated syllable. Another option, reversing the task in some sense, so that listeners respond to the second syllable tone rather the first, would also make a useful comparison, both for the study of anticipatory and carry-over coarticulatory effects and their role in perception. A different task altogether, reiterant speech, might also be profitable to try, although the length of the conversational utterances would be problematic.

The original formulation of the perception experiment included a “noise” condition substituting white noise for the first syllable’s duration to measure the information about first syllable tone present in the second syllable context itself. This would provide a measure of the extent to which listeners can use the second syllable (and coarticulatory effects present in the second syllable), and under what conditions. Because of constraints on the duration of the experiment and its complexity, this condition was deferred to a future study.

For the present perception experiment, one part of the difference between the speech styles (careful vs. conversational) was that careful items were drawn from recordings in isolation, while the conversational items were excised from within whole utterances. This means that there are many important differences in the context of the words. The conversational items happened in the context of a particular position within an utterance, under uncontrolled prosodic circumstances (including questions, discontinuities, fillers, lengthened syllables), potentially in a variety of different functional contexts, and these differences represent part of the measured contrast between the speech styles. A future study should present these full utterances in the careful speech task as well, to be read, in order to control for the influence of many of those differences, in order to understand their role more fully.

The dictation experiment revealed that removal of pitch information does change the distribution of errors with respect to understood morphemes, segments perceived, and tones perceived, but that this change is moderate and uniform across these different units. A natural follow-up study will consider more fully *which* morphemes, segments, and tones listeners have difficulty with, and which are most effected by the manipulation of pitch. It may be that certain prosodic environments make pitch more or less vital to comprehension generally, and the predictability and frequency of units in their context is certainly likely to play a role. The perception of isolated words will also depend on the statistical properties of the Mandarin lexicon, such that some tones may be fully specified by the segmental information (there may be only one real word tone combination for a particular combination of segments in a two-syllable word), and an informational account of listener responses should be undertaken here as well.

In addition to these expansions on the existing experiments, an acoustic survey of Mandarin tones in conversation as compared to careful speech is clearly a priority. Once we understand what happens to the established cues to tone, and their relative utility for distinguishing the tone categories in the two speech styles, clearer predictions will be evident about when listeners will struggle in perception, and about when the articulation of cues makes little difference in perception. The present dissertation is merely a first step in the elucidation of these important topics.

4.6 Conclusion

Mandarin tone categories are represented and perceived using large sets of acoustic cues (not exclusively, or even mainly, pitch), which are affected dramatically by speech style. Conversational tones are distinguished in a far more nuanced and complex way than those in careful speech, but a way which nonetheless provides plentiful information to tone categories on the whole.

The perception experiment highlighted the importance of spectral detail, preserved in the "whisper" stimuli, over the one-dimensional pitch signal provided in the "hum" stimuli. Even in the careful speech words, the "hum" condition was the worst for tone perception.

In the dictation experiment, a 1/3 increase in errors accompanied the removal of overt pitch information from conversational utterances, but the overall number of items with error rates close to zero remained high. When listeners are engaged in the comprehension of conversational speech, pitch information is helpful in a certain portion of the utterances confronting them, but is not vital generally. Rich information is still present in the acoustic detail of the speech.

Conversation, then, challenges the ubiquitous claim that tone is vitally centered on pitch, and presents it as a diversely determined speech category or sound entity. Research aiming to understand the human language faculty would do well to connect theory from highly-controlled traditional research to our most ordinary of social communications.

References

Abramson, A. 1972. Tonal experiments with whispered Thai. In Valdman, editor, *Papers in linguistics and phonetics to the memory of Pierre Delattre*, pages 31–44. Mouton, The Hague.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Berry, J. 2009. Tone space reduction in Mandarin Chinese. *Journal of the Acoustical Society of America*, 125(4):2571.

Bissiri, M. P., Zellers, M., and Ding, H. 2014. Perception of glottalization in varying pitch contexts in mandarin chinese. In *Proceedings of Speech Prosody 2014, Dublin, Ireland*, pages 633–637.

Blicher, D. L., Diehl, R. L., and Cohen, L. B. 1990. Effects of syllable duration on the perception of the mandarin tone 2/tone 3 distinction: Evidence of auditory enhancement. *Journal of Phonetics*, 18(1):37–49.

Bloomfield, L. 1933. *Language*. University of Chicago Press.

Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, volume 17.

Boersma, P. and Weenink, D. 2015. Praat: doing phonetics by computer (version 5.4.08). [Computer program], retrieved 24 Mar 2015 from `http://www.praat.org/`.

Brenner, D. 2013. The acoustics of Mandarin tones in careful and conversational speech. San Francisco. Acoustical Society of America.

Cabrera, L., Tsao, F.-M., Gnansia, D., Bertoncini, J., and Lorenzi, C. 2014. The role of spectro-temporal fine structure cues in lexical-tone discrimination for french and mandarin listeners. *The Journal of the Acoustical Society of America*, 136(2):877–882.

Canavan, A. and Zipperlen, G. 1996. CALLHOME Mandarin Chinese speech.

Cao, J. 1992. On neutral-tone syllables in Mandarin Chinese. *Canadian Acoustics*, 20(3):49–50.

CASS, 中国社会科学院, 语言研究所, 词典编辑室 Chinese Academy of Social Sciences, L. R. D. D. E. O. 2012. 现代汉语词典 Xiandai Hanyu Cidian (Contemporary Mandarin Dictionary). 商務印书馆 Shangwu Yinshu Guan (Commercial Printing), Beijing, 6 edition.

Chandrasekaran, B., Sampath, P. D., and Wong, P. C. 2010. Individual variability in cue-weighting and lexical tone learning. The Journal of the Acoustical Society of America, 128(1):456--465.

Chang, C. and Yao, Y. 2007. Tone production in whispered mandarin. In Proceedings of the 16th International Congress of Phonetic Sciences, pages 326 -- 329, Saarbrücken.

Chao, Y. 1930. A system of tone letters. Le Maitre Phonetique, 45:24--27.

Chao, Y.-R. 1933. Tone and intonation in chinese. 中央研究院歷史語言研究所集刊, pages 121--134.

Chao, Y. R. 1965. A grammar of spoken Chinese. Univ of California Press.

Chen, M. Y. 2000. Tone sandhi: Patterns across Chinese dialects, volume 92. Cambridge University Press.

Chen, M. Y. 2012. Tone rule typology. In Annual Meeting of the Berkeley Linguistics Society, volume 18.

Chen, Y. and Xu, Y. 2006. Production of weak elements in speech: evidence from f0 patterns of neutral tone in standard Chinese. Phonetica, 63:47 -- 75.

Chen, Z. and Lee, K.-F. 2000. A new statistical approach to chinese pinyin input. In Proceedings of the 38th annual meeting of the Association for Computational Linguistics, pages 241--247. Association for Computational Linguistics.

Cheng, C. 2012. Mechanism of extreme phonetic reduction: evidence from Taiwan Mandarin. PhD thesis, University College London.

Cheng, C., Chen, J.-Y., and Gubian, M. 2013. Are mandarin sandhi tone 3 and tone 2 the same or different? the results of functional data analysis. Sponsors: National Science Council, Executive Yuan, ROC Institute of Linguistics, Academia Sinica NCCU Office of Research and Development, page 296.

Cheng, C. and Xu, Y. 2009. Extreme reductions: contraction of disyllables into monosyllables in Taiwan Mandarin. In Interspeech, Brighton, U.K.

Cheng, C. and Xu, Y. 2013. Articulatory limit and extreme segmental reduction in Taiwan Mandarin. Journal of the Acoustical Society of America, 134(6):4481--4495.

Cheng, C. and Xu, Y. 2014. Mechanism of disyllabic tonal reduction in taiwan mandarin. Language and Speech, page 0023830914543286.

Cheng, C., Xu, Y., and Gubian, M. 2010. Exploring the mechanism of tonal contraction in Taiwan Mandarin. In Proceedings of Interspeech, Makuhari, Japan.

Cheng, C., Xu, Y., and Prom-on, S. 2011. Modeling extreme tonal reduction in Taiwan Mandarin based on target approximation. In Proceedings of the International Congress of Phonetic Sciences 17, August 17-21, Hong Kong.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. 1952. Some experiments on the perception of synthetic speech sounds. Journal of the Acoustical Society of America, 24(6):597--606.

Cooper, F. S., Liberman, A. M., and Borst, J. M. 1951. The interconversion of audible and visible patterns as a basis for research in the perception of speech. Proceedings of the National Academy of Science, 37:318 -- 325.

Cutler, A. and Chen, H. C. 1997. Lexical tone in Cantonese spoken-word processing. Perception & Psychophysics, 59(2):165 -- 179.

Day, W. H. E. 1984. Properties of Levenshtein metrics on sequences. Bulletin of Mathematical Biology, 46:327--332.

Dryer, M. S. and Haspelmath, M., editors 2013. WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Duanmu, S. 2007. The Phonology of Standard Chinese. Oxford University Press, Oxford.

Ernestus, M., Baayen, H., and Schreuder, R. 2002. The recognition of reduced word forms. Brain and language, 81(1):162--173.

Ernestus, M. and Warner, N. 2011. An introduction to reduced pronunciation variants. Journal of Phonetics, 39:253 -- 260.

Fant, G. 1973. Speech sounds and features. The MIT Press.

Forster, K. I. and Davis, C. 1984. Repetition priming and frequency attenuation in lexical access. Journal of experimental psychology: Learning, Memory, and Cognition, 10(4):680.

Fowler, C. A. 1984. Segmentation of coarticulated speech in perception. Perception & Psychophysics, 36(4):359--368.

Fromkin, V. 1978. Tone: a linguistic survey. Academic Press.

Gahl, S. 2008. Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. Language, 84(3):474--496.

Greenberg, S. 1999. Speaking in shorthand---a syllable-centric perspective for understanding pronunciation variation. Speech Communication, 29:159 -- 176.

Halle, M. and Stevens, K. 1962. Speech recognition: A model and a program for research. IRE Transactions on Information Theory, 8(2):155--159.

Holt, L. L. and Lotto, A. J. 2006. Cue weighting in auditory categorization: implications for first and second language acquisition. Journal of the acoustical society of America, 119(5):3059--3071.

Howie, J. M. 1976. Acoustical studies of Mandarin vowels and tones, volume VI of Princeton--Cambridge studies in Chinese linguistics. Cambridge University Press, Cambridge.

Hu, F. 2004. Tonal effect on vowel articulation in a tone language. In International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages.

Huang, S., Bian, X., Wu, G., and McLemore, C. 1997. LDC Mandarin lexicon.

Johnson, K. 2004. Massive reduction in conversational american english. In Spontaneous speech: data and analysis. Proceedings of the 1st session of the 10th international symposium, pages 29--54. Citeseer.

Kent, R. D. and Minifie, F. D. 1977. Coarticulation in recent speech production models. Journal of Phonetics, 5(2):115--133.

Kiriloff, C. 1969. On the auditory perception of tones in mandarin. Phonetica, 20(2-4):63--67.

Klatt, D. H. 1973. Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. The Journal of the Acoustical Society of America, 53(1):8--16.

Kong, Y.-Y. and Zeng, F.-G. 2006. Temporal and spectral cues in Mandarin tone recognition. The Journal of the Acoustical Society of America, 120(5):2830--2840.

Kopp, G. and Green, H. 1946. Basic phonetic principles of visible speech. Journal of the Acoustical Society of America, 18(1):74 -- 89.

Kuperman, V., Ernestus, M., and Baayen, H. 2008. Frequency distributions of uniphones, diphones, and triphones in spontaneous speech. The Journal of the Acoustical Society of America, 124(6):3897--3908.

Ladefoged, P. and Broadbent, D. 1957. Information conveyed by vowels. Journal of the Acoustical Society of America, 29(1):98 -- 104.

Levenshtein, V. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics -- Doklady, 10:707--710.

Lewis, M. P., Simons, G. F., and Fennig, C. D., editors 2015. Ethnologue: Languages of the World. SIL International, Online resource, http://www.ethnologue.com, 18 edition.

Li, B. and Guo, Y. 2012. Mandarin tone contrast in whisper. In Proc. 3rd Int. Symp. on Tonal Aspects of Lang., Nanjing, page 84.

Lin, H. 2006. Mandarin neutral tone as a phonologically low tone. Journal of Chinese Language and Computing, 16(2):121--134.

Lin, Y.-H. 2007. The Sounds of Chinese. Cambridge University Press, Cambridge.

Lindstrom, M. J. and Bates, D. M. 1988. Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association, 83(404):1014--1022.

Liu, S. and Samuel, A. G. 2004. Perception of Mandarin lexical tones when f_0 information is neutralized. Language and Speech, 47(2):109 -- 138.

Maddieson, I. 2011. Tone. In Dryer, M. S. and Haspelmath, M., editors, The World Atlas of Language Structures Online, chapter 13. Max Planck Digital Library, Munich.

Miller, J. L., Grosjean, F., and Lomanto, C. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. Phonetica, 41(4):215--225.

Moreno, P. J. and Stern, R. M. 1994. Sources of degradation of speech recognition in the telephone network. In Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on, volume 1, pages I--109. IEEE.

Murthy, S., Rallon, M., Mishra, K., Prakash, S., and Dutta, K. 1979. A guide to pinyin. China Report, 15(3):35--39.

Öhman, S. E. 1966. Coarticulation in vcv utterances: Spectrographic measurements. The Journal of the Acoustical Society of America, 39(1):151--168.

O'Shaughnessy, D. 1988. Linear predictive coding. IEEE Potentials, 7(1):29 -- 32.

Pluymaekers, M., Ernestus, M., and Baayen, R. 2005a. Articulatory planning is continuous and sensitive to informational redundancy. Phonetica, 62:146--159.

Pluymaekers, M., Ernestus, M., and Baayen, R. H. 2005b. Lexical frequency and acoustic reduction in spoken dutch. The Journal of the Acoustical Society of America, 118(4):2561--2569.

Potter, R. K. 1945. Visible patterns of sound. Science, 102(2654):463 --470.

Premaratne, P. 2014. Sign languages of the world. In Human Computer Interaction Using Hand Gestures, pages 145--169. Springer.

Reynolds, D. et al. 1995. Large population speaker identification using clean and telephone speech. Signal Processing Letters, IEEE, 2(3):46--48.

Schertz, J. 2014. The structure and plasticity of phonetic categories across languages and modalities. PhD thesis, University of Arizona.

Schneider, W., Eschman, A., and Zuccolotto, A. 2012. E-Prime Reference Guide. Psychology Software Tools Inc., Pittsburgh.

Schuppler, B., van Dommelen, W. A., Koreman, J., and Ernestus, M. 2012. How linguistic and probabilistic properties of a word affect the realization of its final /t/: Studies at the phonemic and sub-phonemic level. Journal of Phonetics, 40(4):595--607.

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

Shih, C. 1997. Mandarin third tone sandhi and prosodic structure. Linguistic Models, 20:81--124.

Steinberg, J. C. 1934. Application of sound measuring instruments to the study of phonetic problems. Journal of the Acoustical Society of America, 6:16 -- 24.

Story, B. H. and Bunton, K. 2010. Relation of vocal tract shape, formant transitions, and stop consonant identification. Journal of Speech, Language, and Hearing Research, 53(6):1514--1528.

Surendran, D. and Levow, G. 2003. The functional load of tone in Mandarin is as high as that of vowels. In In Proceedings of the International Conference on Speech Prosody 2004, pages 99--102.

Surendran, D. and Niyogi, P. 2003. Measuring the usefulness (functional load) of phonological contrasts. Technical report, University of Chicago.

Taft, M. and Chen, H.-C. 1992. Judging homophony in Chinese: The influence of tones. Advances in psychology, 90:151--172.

Tseng, C.-C. 2004a. Prosodic properties of intonation in two major varieties of Mandarin Chinese: Mainland China vs. Taiwan. In International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages.

Tseng, S. 2004b. Spontaneous Mandarin production: results of a corpus-based study. In Proceedings of the international symposium on Chinese spoken language processing.

Tucker, B. V. and Warner, N. 2007. Inhibition of processing due to reduction of the American English flap. In Proceedings of the International Congress of Phonetic Sciences, volume 16, pages 1949 -- 1952, Saarbrücken.

Wang, W. and Li, K.-P. 1967. Tone 3 in pekinese. Journal of speech and hearing research, 10(3):629.

Warner, N. 2011. The Blackwell Companion to Phonology, chapter Reduction (Chap. 79), pages 1866--91. Blackwell.

Warner, N. 2012. Methods for studying spontaneous speech. In Cohn, A., Fougeron, C., and Huffman, M., editors, Handbook of laboratory phonology. Mouton de Gruyter.

Warner, N., Brenner, D., Schertz, J., Carnie, A., Fisher, M., and Hammond, M. 2015. The aerodynamic puzzle of nasalized fricatives: Aerodynamic and perceptual evidence from scottish gaelic. Laboratory Phonology, 6(2):197--241.

Warner, N. and Tucker, B. V. 2011. Phonetic variability of stops and flaps in spontaneous and careful speech. Journal of the Acoustical Society of America, 130(3):1606 -- 1617.

Watkins, J. W. 2001. Burmese. Journal of the International Phonetic Association, 31(02):291--295.

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

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

Xu, Y. 2004. Transmitting tone and intonation simultaneously --- the parallel encoding and target approximation (PENTA) model. In Proceedings of the international symposium on tonal aspects of languages: with emphasis on tone languages, Beijing, P.R. China.

Yip, M. 2002. Tone. Cambridge University Press, Cambridge, MA.

Yu, K. M. 2010. Laryngealization and features for chinese tonal recognition. In Proceedings of Interspeech, pages 1529--1532.

Zeshan, U. 2008. Roots, leaves and branches--the typology of sign languages. Sign languages: spinning and unraveling the past, present and future, 45:671--695.

Zhao, Y. and Jurafsky, D. 2009. The effect of lexical frequency and lombard reflex on tone hyperarticulation. Journal of Phonetics, 37:231--247.

Appendices

Appendix A

Supplementary Dictionary Entries

⟨披萨⟩	pi1 sa4	1 4	pi sa
⟨物⟩	wu4	4	wu
⟨腹⟩	fu4	4	fu
⟨谏⟩	jian4	4	jyaN
⟨谊⟩	yi4	4	i
⟨萨⟩	sa4	4	sa
⟨苑⟩	yuan4	4	WEn
⟨裕⟩	yu4	4	U
⟨域⟩	yu4	4	U
⟨旮⟩	ga1	1	ga
⟨旯⟩	la2	2	la
⟨肖⟩	xiao4	4	xyaw
⟨灸⟩	jiu3	3	jyow
⟨亚⟩	ya4	4	y@
⟨一个⟩	yi1 ge0	1 0	i g
⟨一个⟩	yi1 ge4	1 0	i g
⟨故事⟩	gu4 shi0	4 0	gu S%
⟨北风⟩	bei3 feng1	3 1	bey f&N
⟨网上⟩	wang3 shang4	3 4	waN SaN
⟨一点点⟩	yi1 dian3 dian3	1 2 3	i dyEn dyEn

⟨看不到⟩	kan4 bu2 dao4	4 2 4	k@n bu daw
⟨告诉⟩	gao4 su0	4 0	gaw su
⟨这个⟩	zhe4 ge0	4 0	Z& g&
⟨闹吧⟩	nao4 ba1	4 1	naw ba
⟨静吧⟩	jing4 ba1	4 1	jiN ba
⟨室友⟩	shi4 you3	4 3	S% yow
⟨真的⟩	zhen1 de0	1 0	Z&n d&
⟨个⟩	ge0	0	g&
⟨手机⟩	shou3 ji1	3 1	Sow ji
⟨一遍⟩	yi1 bian4	1 4	i byEn
⟨因为⟩	yin1 wei4	1 4	yin wey
⟨编程⟩	bian1 cheng2	1 2	byEn C&N
⟨干什么⟩	gan4 shen2 me0	4 2 0	g@n S&n m&