University of Calgary

PRISM Repository **https://prism.ucalgary.ca**

The Vault Open Theses and Dissertations

2017

An exemplar-based model of intonation perception of statements and questions in English, Cantonese, and Mandarin

Chow, Una Yu Po

Chow, U. Y. (2017). An exemplar-based model of intonation perception of statements and questions in English, Cantonese, and Mandarin (Master's thesis, University of Calgary, Calgary, Canada). Retrieved from https://prism.ucalgary.ca. doi:10.11575/PRISM/24881

http://hdl.handle.net/11023/3819

Downloaded from PRISM Repository, University of Calgary

UNIVERSITY OF CALGARY

An exemplar-based model of intonation perception of statements and questions in English, Cantonese, and Mandarin

by

Una Yu Po Chow

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF ARTS

GRADUATE PROGRAM IN LINGUISTICS

CALGARY, ALBERTA

APRIL, 2017

© Una Yu Po Chow 2017

Abstract

To better understand how humans can perceive intonation from speech that includes natural variability, this study investigated whether exemplar theory could account for native listeners' categorization of sentence intonation in English, Cantonese, and Mandarin. In each language, twenty native listeners classified gated utterances of statements and echo questions produced by native speakers. Then a computational model simulated the classification of these utterances, using an exemplar-based process of categorization that relied on F0 only.

The computational model correctly classified these sentences above chance without normalizing F0 by speaker. Compared to the human listeners, the model was similarly sensitive to the cross-linguistic differences in the cues for questions, but performed worse when these cues were (partly) excluded from the utterances. These results suggest that human listeners store whole intonation patterns in memory and use additional acoustic information, along with F0, to categorize new statements and questions, in accordance with exemplar theory principles.

Preface

Parts of this research have been reported in the following conference papers:

Chow, U. Y., & Winters, S. J. (2015). Exemplar-based classification of statements and questions in Cantonese. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. Paper number 0987.1-5.

Chow, U. Y., & Winters, S. J. (2016). Perception of intonation in Cantonese: Native listeners versus an exemplar-based model. *Proceedings of the 2016 Annual Conference of the Canadian Linguistic Association*.

Chow, U. Y., & Winters, S. J. (2016). Perception of statement and question intonation: Cantonese versus Mandarin. *Proceedings of the 16th Australasian International Conference on Speech Science and Technology*, 13-16.

Chow, U. Y., & Winters, S. J. (2016). The role of the final tone in signaling statements and questions in Mandarin. *Proceedings of the 5th International Symposium on Tonal Aspects of Languages*, 167-171.

My co-author permits any content from these papers to be included in this thesis (see Appendix C).

The University of Calgary Conjoint Faculties Research Ethics Board had approved this research study.

Acknowledgements

This thesis would not be possible without the help of many people. First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Steve Winters, for his firm guidance and extreme patience. I could not have conceived this project without his invaluable ideas and his mentorship on the Program for the Undergraduate Research Experience (PURE) project. Prof. Winters enabled me to pursue my long-time goal to be a researcher of intonation. Prof. Darin Flynn also encouraged me to pursue my study in intonation. From the first three books that he recommended—Gussenhoven, Ladd, and Wells—I was determined to follow my dream. Thank you, Prof. Flynn!

My sincere appreciation goes to my thesis examination committee—Prof. Winters, Prof. Flynn, and Prof. Penny Pexman. I want to thank Prof. Pexman especially for her interest in my work. My appreciation extends to Prof. Susanne Carroll, as well, for being the neutral chair for my examination and for providing me with her wise and practical advice throughout my MA study.

I would like to thank Riley Steel, Lindsay Samuel, and Meghan Kruger for helping me annotate part of the English production data in Praat. I would also like to thank Lisa Wong, Danny Chow, and Mingyu Qiu for allowing me to record their productions of the Cantonese and Mandarin tones. I would like to thank my native speakers and listeners for their interest and participation in my research study.

Graduate school would not be complete without fun! Many thanks go to Joey Windsor, Danica MacDonald, Svitlana Winters, Jacqueline Jones, Sarah Greer, and Kelly Burkinshaw for the happy memories of the beyond-academic life (the volunteer and social activities for United Way, A Higher Clause, LLC, GSA, etc.).

I would like to thank the Social Sciences and Humanities Research Council (SSHRC) of Canada, the Province of Alberta, the Faculty of Arts, the Faculty of Graduate Studies, and most of all, the Linguistics Graduate Program for funding me through my MA study.

I am grateful to my physicians, especially Dr. Paterson, Dr. Lui, Dr. Pickering, Dr. McMeekin, Dr. Montgomery, and Dr. Cho, who worked together to save my life and to keep me going, enabling me to pursue my goal and to, hopefully, make a positive difference in people's lives.

Above all, I would like to thank my family for encouraging me when I was struggling and for tolerating me when I was stressed and overwhelmed. I would like to thank Angela and my mom, especially, for helping me with groceries and for cooking homemade soups for me when I was too busy, tired, or ill. 多謝您!

To my father, who was always proud of me for being just me

Table of Contents

Abstract ii
Preface iii
Acknowledgements iv
Dedication vi
Table of Contents vii
List of Tables xi
List of Figures xii
List of Symbols, Abbreviations xvi
Epigraph xviii

CHAPTER ONE: INTRODUCTION 1
1.1 Variability in Speech Production 1
1.2 Exemplar Theory of Speech Perception 3
1.2.1 Exemplar-based Perception of Vowels and Words 3
1.2.2 Exemplar-based Perception of Intonation 5
1.3 Research Questions and Methodology 6
1.3.1 Research Scope 8
1.4 Summary 9

CHAPTER TWO: PROSODY IN ENGLISH, CANTONESE, AND MANDARIN 10
2.1 About the Languages 10
2.2 Prosody 11
2.3 Syllable Structure 12
2.4 Word Stress 15
2.5 Lexical Tones 15
2.6 Statement and Question Intonation 18
2.6.1 Autosegmental-Metrical Theory 18
2.6.2 English, Cantonese, and Mandarin Intonation Patterns 20
2.7 Summary 25

CHAPTER THREE: PROPOSED EXEMPLAR-BASED MODEL OF INTONATION PERCEPTION 26
3.1 Overview of the Model 26
3.2 Exemplar-based Process of Categorization 27
3.3 Auditory Properties for Similarity Calculation 29
3.4 Training and Testing Using Cross-Validation 32
3.5 Summary 34

CHAPTER FOUR: PRODUCTION STUDY 35
4.1 Goals 35
4.2 Methods 35
4.2.1 Participants 35
4.2.2 Stimuli 36
4.2.3 Procedure 41
4.2.4 Acoustic Analysis 43
4.2.5 Exemplar-based Classification 43
4.3 Results 46
4.3.1 Acoustic Analysis 46
4.3.2 Exemplar-based Classification 52
4.4 Discussion 55
4.5 Summary 57

CHAPTER FIVE: PERCEPTION STUDY 58
5.1 Goals 58
5.2 Methods 58
5.2.1 Participants 59
5.2.2 Stimuli 60
5.2.3 Identification Task 67
5.2.4 Procedure 68
5.2.5 Statistical Analysis 71

5.3 Results 72
5.3.1 Perceptual Sensitivity: Across Languages 73
5.3.2 Perceptual Sensitivity: Effect of Final Stress/Tone 75
5.3.3 Perceptual Sensitivity: Between Tone Languages 77
5.3.4 Response Bias: Across Languages 78
5.3.5 Response Bias: Effect of Final Stress/Tone 80
5.3.6 Response Bias: Between Tone Languages 81
5.3.7 Reaction Time: Across Languages 82
5.3.8 Reaction Time: Between Sentence Types 83
5.4 Discussion 84
5.4.1 Cross-linguistic Performance 84
5.4.2 Performances Across Stimulus Types 86
5.4.3 Effects of Final Stress/Tone on Performance 88
5.4.4 Listeners' Response Bias 90
5.4.5 Reaction Time to the Intonation Cue 91
5.5 Summary 93

CHAPTER SIX: PERCEPTUAL SIMULATIONS OF THE MODEL 95
6.1 Goals 95
6.2 Methods 95
6.2.1 Simulated Listeners 95
6.2.2 Stimuli 95
6.2.3 Classification Task 98
6.2.4 Procedure 100
6.2.5 Statistical Analysis 100
6.3 Results 101
6.3.1 Perceptual Sensitivity: Across Languages 101
6.3.2 Perceptual Sensitivity: Effect of Final Stress/Tone 104
6.3.3 Response Bias: Across Languages 110
6.4 Discussion 112
6.4.1 The Computer Model's Performance 112

6.4.2 The Computer Model versus Human Listeners 113
6.4.3 Cross-linguistic Performance 116
6.4.4 Effects of Final Stress/Tone on Intonation Perception 117
6.4.5 Listeners' Response Bias 120
6.4.6 The Human Listeners' Perception of Intonation 121
6.5 Summary 123

CHAPTER SEVEN: TOWARDS A GENERALIZED INTONATION PERCEPTION MODEL 124
7.1 The Kernel Model 124
7.2 Fine-tuning the Model 125
7.3 Additional Mechanisms for the Model 136
7.4 Considerations for a Generalized Model 138
7.5 Summary 141

CHAPTER EIGHT: CONCLUSION 143
8.1 Findings 143
8.2 Contribution 146
8.3 Limitations 146
8.4 Future Directions 147

REFERENCES 149

APPENDIX A: STIMULI 160
A.1 English Stimuli 160
A.2 Cantonese Stimuli 163
A.3 Mandarin Stimuli 171

APPENDIX B: BACKGROUND QUESTIONNAIRE 179

APPENDIX C: LETTER OF COPYRIGHT PERMISSION 180

List of Tables

Table 2.1 Cantonese tones 16

Table 2.2 Mandarin tones 17

Table 4.1 Demographics of the speakers in the production study 35

Table 4.2 Mapping between Mandarin and Cantonese tones 39

Table 4.3 Mandarin and Cantonese target syllables and tones, initially and finally 40

Table 4.4 Coefficients and *p*-values of a logistic regression on SpC 54

Table 4.5 Coefficients and *p*-values of a logistic regression on MpC 55

Table 5.1 Demographics of the listeners in the perception study 59

Table 5.2 Stimulus types 60

Table 5.3 The stimulus type(s) and number of trials presented in each part 67

Table 5.4 Ten listener orders, generated from five random orders of the stimuli 68

Table 5.5 Application of signal detection theory to ‘statement’ and ‘question’ responses 71

Table 5.6 Interaction between language and stimulus type on d' 74

Table 5.7 Interaction between language and stimulus type on ß 80

Table 5.8 Interaction between language and stimulus type on normalized RT 83

Table 6.1 The stimulus type and number of trials presented in each simulation 99

Table 6.2 Stimulus sets used for runs #1 and #2 of the 2-fold cross-validation 99

Table 6.3 Interaction between language and stimulus type on d', model only 103

List of Figures

Figure 2.1 English syllable structure 13

Figure 2.2 Cantonese syllable structure 13

Figure 2.3 Mandarin syllable structure 14

Figure 2.4 F0 contours of a male speaker's production of *ji* with the six Cantonese tones 16

Figure 2.5 F0 contours of a female speaker's production of *yi* with the four Mandarin tones 18

Figure 2.6 Intonation patterns of an English statement and echo question produced by a male speaker 20

Figure 2.7 Intonation contours of a Cantonese statement and echo question produced by a female speaker: 'Wong Ji is not on time' 22

Figure 2.8 Intonation contours of a Cantonese statement and echo question produced by a female speaker: 'Wong Ji teaches history' 23

Figure 2.9 Intonation contours of a Mandarin statement and echo question produced by a female speaker: 'Wang Wu watches TV' 24

Figure 3.1 F0 values at eleven equidistant time points of an utterance of 'films' 29

Figure 3.2 Annotation of an English question produced by a female speaker 30

Figure 3.3 'PointProcess' object in Praat for the utterance in Figure 3.2 30

Figure 3.4 F0 contour of the utterance in Figure 3.2 before and after applying interpolation 31

Figure 3.5 Eleven equidistant time points of the pitch contour in Figure 3.4(b) 31

Figure 3.6 Three-fold cross-validation 33

Figure 4.1 English, Cantonese, and Mandarin dialogues presented to the speakers... 41
Figure 4.2 Mean F0 contours by speaker gender and sentence type in English, Cantonese, and Mandarin 47
Figure 4.3 Mean F0 contours by final stress and sentence type in English 49
Figure 4.4 Mean F0 contours by final tone and sentence type in Cantonese..... 50
Figure 4.5 Mean F0 contours by final tone and sentence type in Mandarin..... 51
Figure 4.6 Single-point Classification versus Multi-point Classification..... 53
Figure 5.1 A marked textgrid for segmenting into the five stimulus types 62
Figure 5.2 F0 contours of stimulus types Whole, Last2, and Last..... 63
Figure 5.3 F0 contours of stimulus types NoLast and First 64
Figure 5.4 Final stress or tone in English, Cantonese, and Mandarin 66
Figure 5.5 Numerical keys corresponding to the gradient and categorical responses..... 69
Figure 5.6 Interaction between language and stimulus type on d' 74
Figure 5.7 Interaction between stimulus type and stress on d' for English..... 75
Figure 5.8 Interaction between stimulus type and tone on d' for Cantonese 76
Figure 5.9 Interaction between stimulus type and tone on d' for Mandarin 77
Figure 5.10 Interaction among language, stimulus type, and tone on d'..... 78
Figure 5.11 Interaction between language and stimulus type on ß..... 79
Figure 5.12 Interaction between stimulus type and tone on ß for Mandarin 81
Figure 5.13 Interaction between language and stimulus type on normalized RT..... 82
Figure 5.14 Interaction between stimulus type and sentence type on normalized RT 84

Figure 6.1 F0 ranges of the speakers' production of the first two syllables 97
Figure 6.2 Interaction among listener type, language, and stimulus type on d' 102
Figure 6.3 Interaction between language and stimulus type on d', model only 103
Figure 6.4 Interaction among listener type, stimulus type, and stress on d' for English 104
Figure 6.5 Interaction between stimulus type and stress on d' for English, model only 105
Figure 6.6 Interaction among listener type, stimulus type, and tone on d' for Cantonese 106
Figure 6.7 Interaction between stimulus type and tone on d' for Cantonese, model only 107
Figure 6.8 Interaction among listener type, stimulus type, and tone on d' for Mandarin 108
Figure 6.9 Interaction between stimulus type and tone on d' for Mandarin, model only 109
Figure 6.10 Interaction among listener type, language, and stimulus type on ß 111
Figure 7.1 Variation in the timing of the nuclear accent in two different productions of the same question 127
Figure 7.2 English question rise at a final stressed syllable 127
Figure 7.3 Misalignment of the question rise between a token (bottom) and an exemplar (top) in static time comparison 128
Figure 7.4 Dynamic time alignment process of an exemplar (top, red) with a token (bottom, blue), using three window lengths 130

Figure 7.5 Dynamic time alignment of an exemplar (top, red) with a token (bottom, blue) using five window lengths 131

Figure 7.6 Alignment of two fragments with a whole utterance (question, top; or statement, bottom) through 11 comparisons 133

Figure 7.7 F0 ranges of the Mandarin speakers' production of stimulus type First, averaged over all five blocks 134

Figure 7.8 Intonation cues (red dotted lines) for a Mandarin statement and question: "Wang Wu watches TV" 135

Figure 7.9 Categorization of two tokens of "Wang Wu teaches history" (middle) by sentence-type intonation (top) and final tone (bottom) 140

List of Symbols, Abbreviations

Symbol, Abbreviation	Definition
ANOVA	analysis of variance
ß	beta (a measure of response bias)
C	consonant
d'	d-prime (a measure of perceptual sensitivity)
dB	decibel
F	falling tone
F0	fundamental frequency (perceived as pitch)
F1	first formant
F2	second formant
F3	third formant
H	high tone
H%	high boundary tone (used in ToBI transcription)
H*	high pitch accent (used in ToBI transcription)
H-	high phrase accent (used in ToBI transcription)
Hz	Hertz
L	low tone
L%	low boundary tone (used in ToBI transcription)
L*	low pitch accent (used in ToBI transcription)
L-	low phrase accent (used in ToBI transcription)
L+H*	low tone, followed by a high, stressed tone
L*+H	low, stressed tone, followed by a high tone

maxF0	maximum F0
meanF0	mean F0
minF0	minimum F0
MpC	Multi-point Classification (in exemplar-based modeling)
R	rising tone
RT	reaction time
SpC	Single-point Classification (in exemplar-based modeling)
ToBI	Tones and Break Indices (annotation system for intonation)
Tukey HSD	Tukey's Honest Significant Difference (statistical test)
V	vowel
w	attention weight (in exemplar-based modeling)
$\bar{x}$	sample mean
z	z-score

It does not matter how slowly you go so long as you do not stop.

-- Confucius

bù pà màn, jiù pà zhàn

-- Kǒng Zǐ

不怕慢，就怕站

-- 孔子

Chapter 1: Introduction

1.1 Variability in Speech Production

Speech perception researchers have recognized for several decades that there is considerable acoustic-phonetic variation in the production of the phonemic categories of speech. In a classic example, Peterson and Barney (1952) measured the first formant (F1) and second formant (F2) frequencies of 10 vowels in hVd words (*heed*, *hid*, *head*, *had*, *hod*, *hawed*, *hood*, *who'd*, *hud*, and *heard*), produced by 76 English speakers (33 men, 28 women, and 15 children). The scatterplots of these formant frequencies showed wide and overlapping distributions of the vowels, meaning there was no simple or straightforward acoustic correlate for each vowel feature. Peterson and Barney (1952) also found considerable variation among the different groups of speakers: on average, children produced the highest formant frequencies, women produced the second highest formant frequencies, and men produced the lowest formant frequencies. Physical differences in the vocal tract length of the speakers (Fant, 1970, 1972) could primarily account for this cross-speaker variability.

In addition, researchers have long recognized that phonetic context contributes to within-speaker variability. In an experiment investigating the perception of unvoiced stop consonants, Liberman, Delattre, and Cooper (1952) demonstrated that the perceptual identification of the release burst of a stop consonant is affected by the vowel it precedes in a complicated pattern. To test the influence of release burst frequency on the identification of unvoiced stops (/p/, /t/, and /k/), Liberman et al. (1952) created 84 consonant-vowel stimuli by combining 12 synthesized stop bursts (that had frequencies between 360 and 4320 Hz) with seven vowels, and then presented them twice to 30

listeners. The results of this identification test showed that the perception of the synthesized /p/ and /k/ bursts varied depending on the following vowel. Within 1440-1800 Hz, listeners perceived the synthesized stops preceding /i, e, o, u/ mainly as /p/, and the synthesized stops preceding /ɛ, a, ɔ/ mainly as /k/. These results suggest that vowel context affects the acoustic cues that listeners used to identify consonant stops.

Despite this 'lack of invariance', or the lack of a one-to-one mapping between an acoustic signal and its phonemic category, listeners in general are able to interpret the intended utterance. Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967) observed that the stop consonant /d/ in two different vowel contexts exhibits two different F2 transition cues. When followed by /i/ (i.e., /di/), F2 rises from 2200 to 2600 Hz; when followed by /u/ (i.e., /du/), F2 falls from 1200 to 700 Hz. Apparently, both consonant and vowel information is encoded in the F2 transition. Regardless of the variation in F2 transition cues, listeners perceived the stop as /d/ in both vowel contexts. Speech perception researchers are thus faced with the crucial yet challenging question of determining how humans can successfully identify abstract phonemic and lexical categories from the huge amount of context- and speaker-based variation in speech production.

Elman and McClelland (1986), however, claimed that the lack of invariance in speech is not an actual problem for listeners. "It is precisely the variability in the signal which permits listeners to understand speech in a variety of contexts, and spoken by a variety of speakers" (Elman & McClelland, 1986: 360). What they referred to as 'lawful variability' is not noise in the speech signal but information about sources of variability (Magnuson & Nusbaum, 2007: 403; Nygaard, 2005), which can be predicted. The

implication for speech perception is that listeners require detailed phonetic information in order to analyze the sources of phonetic variation in the speech signal. One theoretical approach to speech perception, which makes such detailed phonetic information its foundation, is exemplar theory.

1.2 Exemplar Theory of Speech Perception

1.2.1 Exemplar-based Perception of Vowels and Words

Originating from psychological models of the categorization of objects, exemplar theories of perception claim that a category is represented by all experienced instances ('exemplars') of the category (Hintzman, 1986, 1988; Nosofsky, 1986, 1988). Johnson (1997) adapted Nosofsky's (1986, 1988) model to speech perception and proposed that listeners store the exemplars of speech that they experience in rich phonetic detail in memory, and associate this information with characteristics of the speakers such as their identity, gender, and social class (Johnson 1997, 2006). Since the phonetic details of these exemplars are not 'normalized', or filtered out of their mental representations, the model holds that listeners can use the inherent variability of exemplars to categorize new instances of speech based on how similar these instances are to the exemplars in memory, without the need for 'speaker normalization' (Johnson, 1997, 2005). For example, information about formant differences among speakers is not abstracted away during speech processing.

To test this hypothesis, Johnson (1997) simulated vowel perception using an exemplar-based model (Nosofsky, 1986, 1988). The test tokens consisted of 10 different (h)Vd words, read by 14 male and 25 female native English speakers five times each.

Each of these word tokens was presented to the model for categorization, while the rest of the tokens served as experienced exemplars in memory. The model calculated similarity between each new word and the exemplars in memory based on the weighted values of their vowel properties (F0, F1, F2, F3, and duration). An annealing algorithm (Johnson, 1997; Masters, 1995) determined the weights that would approximate optimal performance. The best-fitting model correctly categorized these word tokens 80% of the time—a success rate that is comparable to human listener performance on synthesized vowels (Johnson, 1997; Lehiste & Meltzer, 1973; Ryalls & Lieberman, 1982). The model's confusion matrix also significantly correlated with the human listeners' confusion matrix in Peterson and Barney's (1952) vowel identification task, which used a similar list of hVd words. The results of this study demonstrated that an exemplar-based model could in principle account for certain aspects of human vowel perception, such as F1 and F2 variation.

Similarly, Goldinger (1998) simulated word perception using Hintzman's (1986, 1988) MINERVA 2. This model assumes that listeners store experienced instances of a word as detailed 'traces' in memory. Abstraction only occurs during word retrieval. When a new token is heard, its acoustic representation (a 'probe') activates traces that are similar to it. An 'echo', determined by the summed feature values of the activated traces in memory, is then retrieved. In Goldinger (1998), the model simulated the AXB test performed by human listeners in Goldinger's (1996) recognition memory experiment. First, the model stored 20 instances of 144 words that differed in voice and context to simulate prior experience, followed by another 72 of these words that differed in voice only to simulate the study phase. Then, the model discriminated 144 old and new words

from the study phase, presented in new voices. To simulate delay periods between study and test, the model applied decaying cycles to the memory traces in the study phase prior to testing. The simulation results corresponded to the human listener results: single-voice stimuli yielded better performance than multiple-voice stimuli, but this voice effect gradually disappeared over time. As with Johnson (1997), the results of this study demonstrated that, when detailed traces of words are stored in memory, explicit normalization is, in theory, not required for categorizing words.

1.2.2 Exemplar-based Perception of Intonation

Variability in speech occurs beyond the vowel and the word levels, as researchers have observed variation in the realization of tone and intonation as well. For example, Flynn (2003) analyzed the six Cantonese tones as produced by five native speakers of Hong Kong Cantonese and found variation in the pitch height and slope of individual tones due to coarticulation with adjacent tones. Carryover and anticipatory effects altered the onset and offset of the target tones, respectively. In addition, Warren (2005) compared the onsets of high-terminal rises in statements and questions produced by two groups of native New Zealand English speakers: 1) six male and six female teenagers between 16 and 19 years old, and 2) six male and six female adults between 30 and 45 years old. Same-sex dyads from each group produced a variety of sentences that were controlled in the study. They also freely produced a variety of sentences while performing a map task (Brown, Anderson, Shillcock, & Yule, 1984). Warren (2005) found that the teenage group produced more high terminal rises that started at the nuclear syllable (i.e., the last prominent stressed syllable) in the intonational phrase, whereas the mid-age group

produced more rises that started at a post-nuclear syllable. Multiple factors, including the tonal context and the speaker's sociolect, contributed to the variations in these examples.

Assuming that human speech perception draws on the rich phonetic details of speech, an exemplar-based model should be able to account for the categorization of varying intonation contours, as well. To date, only a few studies have investigated the classification of prosodic elements by an exemplar-based model. Walsh, Schweitzer, and Schauffer (2013) simulated the categorization of two pitch accents (H*L and L*H) in German using an exemplar-based model (Johnson, 1997; Nosofsky, 1986). They tested their simulation with five hours of broadcast radio speech data and found that both pitch accents could be successfully categorized (at 30% above chance) using the exemplar-based similarity approach. Church and Schacter (1994) altered the word intonation of pairs of statements and questions between study and test phases of a word recognition task and found that this change affected listeners' performance. Their result suggests that word intonation is stored in implicit memory. Calhoun and Schweitzer (2012) also demonstrated that intonation contours of frequently occurring words or phrases could be stored in memory, with their associated collocations. Since these collocations are shorter than a sentence, it is unknown whether humans can (or do) store intonation patterns for whole sentences in memory.

1.3 Research Questions and Methodology

To investigate whether exemplar theory can account for the categorization of sentence intonation, this research study proposed an exemplar-based model that can categorize statements and echo questions based on intonation alone. The proposed model was tested

on three different languages: English, Cantonese, and Mandarin. These three languages were chosen because each language has a distinct intonation system, which provided three unique cases for testing the proposed exemplar-based model. English is a stress language in which echo questions have a rising pitch at the end of an utterance, Cantonese is a tone language in which echo questions also have a final rising pitch, and Mandarin is both a stress and tone language in which echo questions exhibit a global rise in pitch level.

My research questions were as follows: 1) Can an exemplar-based model correctly classify statements and echo questions in English, Cantonese, and Mandarin, based solely on intonation? 2) If so, how well does this model account for native listeners' perception of sentence intonation? Specifically, what can we learn about the human perception of intonation from the differences that emerge between the human and exemplar-based classification of statements and echo questions? 3) Is the model flexible enough to handle the different intonation patterns in all three languages? What differences in performance emerge for the three languages under study?

To address these research questions, I used the following methodology. 1) I created an exemplar-based computational model that can categorize statements and echo questions based solely on intonation. 2) I conducted a production study in which I recorded statements and echo questions from native speakers of English, Cantonese, and Mandarin. 3) I conducted a perception study in which I tested native listeners' performance in identifying statements and echo questions from stimuli created from the recorded utterances. I used a gating task in order to find out how much intonational information listeners could get from each gated portion of an utterance. 4) I tested the

model's ability to classify the same set of stimuli used in testing the human listeners. 5) I compared the model's performance with the human listeners' performance on the identification task.

My expectations for each research question were as follows. 1) I expected the exemplar-based model to be able to learn how to categorize statements versus echo questions in English, Cantonese, and Mandarin based on their intonation patterns. 2) However, given its lack of prior language experience, I expected this model to perform worse than human listeners in the same identification task, but still better than chance. I anticipated that differences in human and computer performance on this task would provide insight into research question #2. 3) Since English, Cantonese, and Mandarin have distinct intonation patterns for questions, I expected the model to perform differently across all three languages. I anticipated that the differences in the model's performance would reveal which intonation cues for questions were more (or less) salient for the model.

1.3.1 Research Scope

My thesis is grounded on the exemplar theory of speech perception. Although a complete exemplar-theoretic account of speech processing also needs to consider speech production (Kirchner, Moore, & Chen, 2010; Pierrehumbert, 2001; Wedel, 2004), modeling intonation production and the production-perception link is beyond the scope of this thesis. In addition, although hybrid models of abstract and episodic representations are plausible (e.g., Goldinger, 2007), this thesis assumes a purely episodic model (Johnson, 1997). Finally, although the proposed model processes sentence intonation, this

thesis makes no assumption about the units of speech that are stored in memory (e.g., segmental features, segments, words, sentences, or prosodic features); they could be of variable length (Goldinger & Azuma, 2003).

1.4 Summary

This chapter has presented 1) the issue of trying to understand the human perception of speech when faced with massive variability in speech production and 2) an exemplar-theoretic approach to account for vowel and word perception. This introduction has set the stage for my proposal to extend exemplar theory to account for the perception of sentence intonation in English, Cantonese, and Mandarin. The remaining chapters are organized as follows. Chapter 2 describes the prosodic (e.g., stress, tone, and intonation) systems of English, Cantonese, and Mandarin. Chapter 3 introduces my proposed exemplar-based model of intonation perception. Chapter 4, 5, and 6 report the results of the production study, perception study, and simulations on the model, respectively. Chapter 7 discusses the potential of this model. Finally, Chapter 8 concludes with suggestions for further research.

Chapter 2: Prosody in English, Cantonese, and Mandarin

This chapter provides an overview of the languages under study in Section 2.1, along with a description of their prosodic systems.

2.1 About the Languages

English is a West Germanic language of the Indo-European language family. It is the national language of the United Kingdom and one of the two national languages of Canada. English has approximately 339.4 million native speakers, mostly in the United States, United Kingdom, and Canada (Lewis, Simons, & Fennig, 2016). It is also widely spoken in Hong Kong and taught in many ESL schools in China.

Mandarin (or Standard Chinese) is a Chinese language of the Sino-Tibetan language family. It is the national language of China and the educational language of Singapore. Mandarin has approximately 897.1 million native speakers, mostly in mainland China, Taiwan, and Singapore (Lewis et al., 2016). These regional variations are mutually intelligible: Putonghua in China, Guoyu in Taiwan, and Huayu in Singapore. Since Hong Kong's reversion to China in 1997, after a 99-year lease to Britain, the number of Mandarin speakers in Hong Kong has increased, partly due to the increased interaction with mainland China's economy and partly due to the emphasis on Mandarin-language education.

Cantonese (or Yue) is also a Chinese language of the Sino-Tibetan language family. It is the second-most often used language in China, mainly spoken in Guangdong and east Guangxi provinces. It is the *de facto* provincial language in Guangdong Province, Hong Kong, and Macao. Cantonese has approximately 63.0 million native

speakers, mostly in China and Hong Kong, Malaysia, Vietnam, and Macao (Lewis et al., 2016). Many Cantonese speakers are fluent in English or another Chinese language.

Mandarin and Cantonese are mutually unintelligible in their spoken forms, but they share the same Chinese characters in their written forms. Some expressions differ in both spoken and written forms between these two languages (e.g., 係 *hai6* 'yes' in Cantonese versus 是 *shi4* 'yes' in Mandarin). Also, there are two styles of Chinese characters: traditional (e.g., 國 'nation') and simplified (e.g., 国 'nation'). Traditional characters were developed in the 5th century, while simplified characters were adopted in 1956 by the government of the People's Republic of China. Simplified characters are now the official writing style in mainland China and Singapore (as of 1969), while traditional characters remain the official writing style in Taiwan, Hong Kong, and Macau.

2.2 Prosody

Prosody in speech is a pattern of suprasegmental features that are superimposed on segments (consonants and vowels). It includes stress, tone, and intonation. "Stress refers to the rhythmic pattern or relative prominence of syllables in an utterance" (Pierrehumbert & Hirschberg, 1990) and is often characterized by increases in pitch (or fundamental frequency), length (or duration), and loudness (or intensity). Lexical tone is a characteristic pitch pattern that occurs on a syllable of a word and can alter the meaning of that word. Intonation (or speech melody) is the rise and fall in pitch throughout an utterance. A salient acoustic property of all three of these prosodic features is fundamental frequency.

Fundamental frequency (F0), as measured in Hertz (Hz), is the number of times that the vocal folds open and close per second in voicing. According to the myoelastic theory (Reetz & Jongman, 2009: 78), the length and elasticity of the vocal folds affect the speed at which they open and close. First of all, the longer the vocal folds, the slower the voicing cycle. Typically, men's vocal folds are 17-24 mm in length, whereas women's vocal folds are 13-17 mm in length (Raphael, Borden, & Harris, 2011: 70). Consequently, men usually have a lower F0 than women. On average, adult male voices have an F0 of approximately 125 Hz, whereas adult female voices typically have an F0 higher than 200 Hz. Secondly, the thinner or more tense the vocal folds, the faster the voicing cycle. Lengthening–in effect, thinning–and tensing the vocal folds will subsequently raise the F0. Human vocal folds normally vibrate at a rate of 80-500 Hz during speech (Raphael et al., 2011: 30). However, male speakers can produce low tones in Mandarin or Cantonese below 75 Hz, and female speakers can produce high boundary tones in questions above 500 Hz. Thus, gender differences in vocal fold size and shape, along with cross-linguistic differences, create F0 variation in the production of intonation.

2.3 Syllable Structure

"Syllables are necessary units in the organization and production of utterances" (Ladefoged, 1982). A syllable is a combination of one or more segments. The nucleus is the vocalic (vowel) part of a syllable. Since stress and tone are acoustic properties of syllables, I will first describe the syllable structures of English, Cantonese, and Mandarin.

The English syllable consists of an optional onset, followed by an obligatory rhyme. The rhyme consists of an obligatory nucleus, followed by an optional coda, as

shown in Figure 2.1. A syllable can bear stress in English, which I will discuss in Section 2.4.

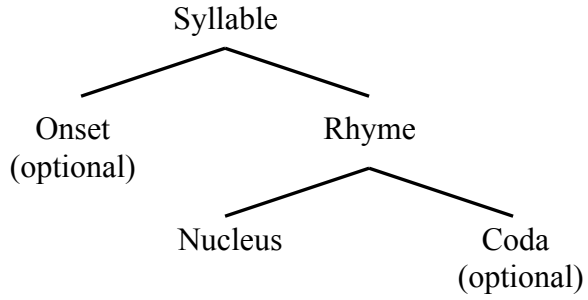


Figure 2.1. English syllable structure.

The onset or coda, if present, can include one or more consonants (e.g., the onset [f] and the coda [lmz] in [ˈfɪlmz] 'films'). The nucleus can be a vowel (e.g., [aɪ] 'I') or a syllabic consonant (e.g., [n̩] in [ˈbʌʔ.n̩] 'button'). Native speakers sometimes differ in their judgment on the syllabification of certain words (e.g., [meɪ.ɹi] or [mɛɪ.i] 'Mary'[1]).

The Cantonese syllable consists of an optional onset, followed by an obligatory rhyme (Bauer & Benedict, 1997: 9)[2]. The rhyme consists of an obligatory nucleus, followed by an optional coda, as shown in Figure 2.2. The Cantonese syllable also carries a lexical tone, which I will discuss in Section 2.5.

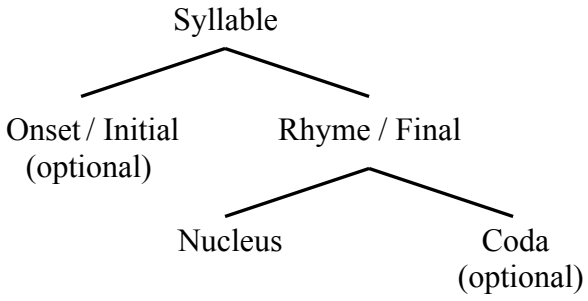


Figure 2.2. Cantonese syllable structure.

[1] Two of my native English-speaking transcribers syllabified 'Mary' differently from each other.
[2] Bauer and Benedict (1997) used the terms 'initial' and 'final'. In this thesis, I referred to 'initial' as 'onset' and 'final' as 'rhyme' to be consistent with the terms used for the English syllable structure.

Not all combinations of onset + nucleus + coda are permissible in this language. For example, a syllabic [m̩] can only be a nucleus in Cantonese if the syllable is one segment long. Examples[3] of Cantonese syllables include *jat* [jɐt] 'one', *si* [siː] 'time', *on* [ɔn] 'press', and *m* [m̩] 'not'.

The Mandarin syllable consists of an optional onset, followed by an obligatory rhyme (Li & Thompson, 1981)[4], as shown in Figure 2.3. The Mandarin syllable also carries a lexical tone, which I will discuss in Section 2.5.

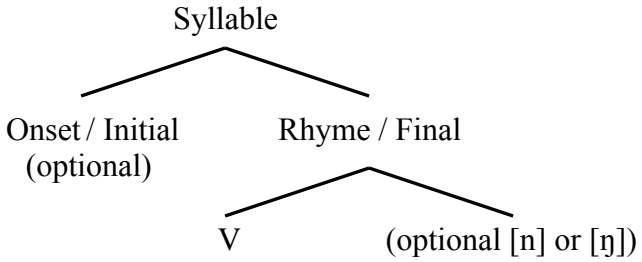


Figure 2.3. Mandarin syllable structure.

Mandarin disallows consonant clusters within a syllable and prefers syllables consisting of only a consonant and a vowel (CV). The only consonants that can appear at the end of a syllable are [n] and [ŋ]. The nucleus can be a monophthong, diphthong, or a triphthong. Examples[5] of Mandarin syllables include *shang* [ʂɑŋ] 'up', *mai* [maɪ] 'buy', and *jiao* [tɕiɑʊ] 'teach'.

[3] Cantonese has many romanization schemes, including Yale, Sidney Lau, and Jyutping. Jyutping (粵拼) was designed by the Linguistic Society of Hong Kong in 1993. Jyutping is represented using only the English alphabet and numbers, and appears frequently in linguistics literature. Therefore, this thesis uses Jyutping to show the Cantonese written examples.

[4] Li and Thompson (1981) used the terms 'initial' and 'final'. In this thesis, I referred to 'initial' as 'onset' and 'final' as 'rhyme' to be consistent with the terms used for the English syllable structure.

[5] Pinyin is the standard romanization system for Mandarin, developed in the 1950s and published by the Chinese government in 1958. This thesis uses Pinyin to show the written Mandarin examples.

2.4 Word Stress

English is a word-based stress language.[6] In English, a stressed syllable is realized with a relatively higher F0, longer duration, and greater intensity than an unstressed syllable (Raphael et al., 2011: 147). According to Fry (1958), F0 is the primary cue for stress in English; however, Beckman (1986) found intensity and duration to be more salient cues for stress in English. Hirst (1983) suggested that, in speech production, duration and intensity are the dominant signals, while in speech perception, F0 is the dominant cue.

English has a stress pattern in which the stress normally falls on the penultimate syllable of polysyllabic words, (e.g., *teacher* [ˈtʰi.tʃɻ]). However, this pattern differs for some words (e.g., *engineer* [ˌɛ̃n.dʒəˈnɪɹ]). Monosyllabic words are usually stressed (e.g., *friend* [ˈfɹɛnd]), but not always, especially when the word is a function word (e.g., *of* [əv]).

2.5 Lexical Tones

Cantonese and Mandarin are tone languages, in which each word bears a lexical tone. These lexical tones may alter a word's meaning. The tonal inventories of these two languages differ.

Cantonese has six contrastive tones, as shown in Table 2.1. Represented using Chao's (1947) 5-scale tonal system, with 1 being the lowest and 5 being the highest point in a speaker's pitch range, the tones are [55], [25], [33], [21], [23], and [22]. Every Cantonese syllable has a specified tone, which is carried by the rhyme. In the Jyutping romanization system, the tone number appears at the end of the written syllable. For

[6] Mandarin also has word stress. However, this thesis focuses on the effects of Mandarin tones on intonation because both tones and intonation use F0 as their primary cue (Yuan, 2011). The discussion of Mandarin stress is beyond the scope of this thesis. See Duanmu (2007).

example, *ji* with a high-level tone (Tone 1) is written as *ji1* (meaning 'doctor'), as shown in Table 2.1.

Table 2.1. Cantonese tones (source: Flynn, 2003).

Tone	Shape	Pitch level	Example (Jyutping)
1	high-level	55	醫 *ji1* 'doctor'
2	high-rise	25	椅 *ji2* 'chair'
3	mid-level	33	意 *ji3* 'meaning'
4	low-fall	21	疑 *ji4* 'to suspect'
5	low-rise	23	耳 *ji5* 'ear'
6	low-level	22	二 *ji6* 'two'

Figure 2.4 shows the tonal contours of the six minimally contrastive words listed in Table 2.1, produced by a male, native Cantonese speaker. The speaker produced these words in isolation using his normal pitch range. The red dotted line indicates where approximately this speaker produced a '2' on Chao's (1947) 5-scale tonal system (e.g., [25], [21], and [22]). Since pitch levels are relative to the speaker's pitch range, a '2' produced by a different speaker could be higher or lower than this speaker's pitch value.

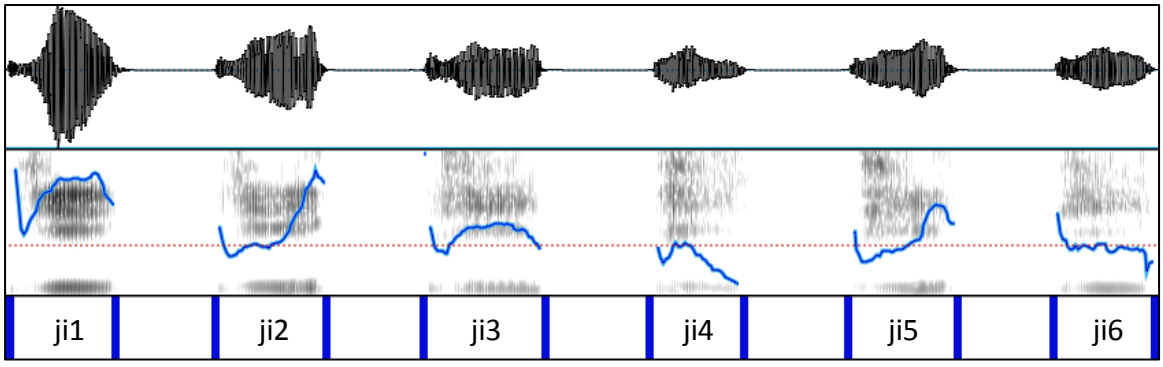


Figure 2.4. F0 contours of a male speaker's production of *ji* with the six Cantonese tones. The pitch setting is 75 to 150 Hz. The red dotted line crosses at 100 Hz.

Mandarin, on the other hand, has four contrastive tones, as shown in Table 2.2. Represented using Chao's (1948) 5-scale tonal system, the tones are [55], [35], [21(4)], and [51]. In addition, there is a neutral tone. This tone is unspecified and assimilates the tone from the specified tone of the immediately preceding syllable. Tone 3 has two allotones: [21] and [214]. The [21] variant occurs only in final stressed syllables (Hartman, 1944). Similar to Cantonese tones, Mandarin tones are carried by the rhyme. The tone number also appears at the end of the written syllable in Pinyin, as the examples in Table 2.2 show.

Table 2.2. Mandarin tones (source: Li & Thompson, 1981).

Tone	Shape	Pitch level	Example (Pinyin)
1	high level	55	医 *yi1* 'doctor'
2	high rising	35	疑 *yi2* 'to suspect'
3	falling rising	21(4)	椅 *yi3* 'chair'
4	high falling	51	意 *yi4* 'meaning'

Figure 2.5 shows the tonal contours of the four minimally contrastive words listed in Table 2.2, produced by a female, native Mandarin speaker. The speaker produced these words in isolation using her normal pitch range. The red dotted line indicates where approximately this speaker produced a '5' on Chao's (1948) 5-scale tonal system (e.g., [55], [35], and [51]). Again, since pitch levels are relative to the speaker's pitch range, a '5' produced by a different speaker could be higher or lower than this speaker's pitch value.

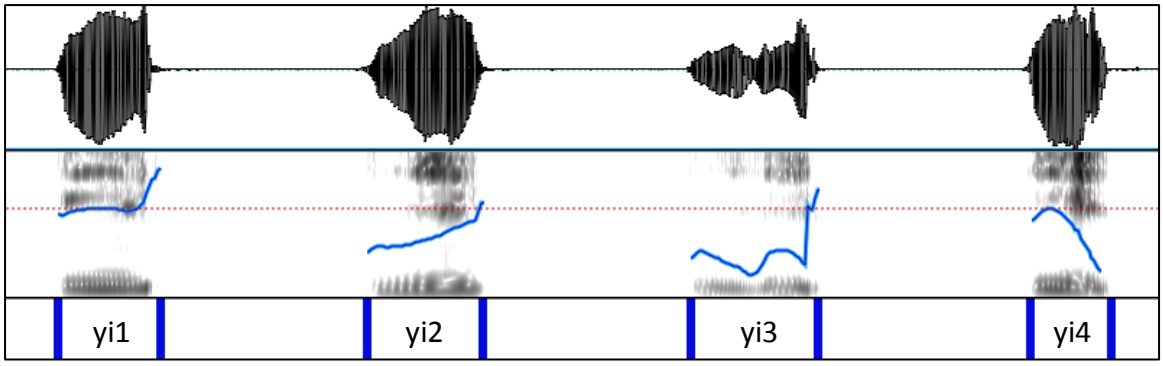


Figure 2.5. F0 contours of a female speaker's production of *yi* with the four Mandarin tones. The pitch setting is 75 to 425 Hz. The red dotted line crosses at 287 Hz.

2.6 Statement and Question Intonation

2.6.1 Autosegmental-Metrical Theory

This thesis describes intonation patterns using the autosegmental-metrical theoretical framework. The autosegmental-metrical theory (Goldsmith, 1976, 1990) claims that suprasegmental features, such as syllable, stress, tone, and intonation are represented phonologically in hierarchical layers, autonomous of the segmental layer. Based on this theory, Pierrehumbert (1980) developed a framework for mapping the phonological categories of English intonation to their phonetic realizations. In this model, intonation is structured in hierarchical units. The largest unit, the 'intonational phrase', is a sentence or a phrase that is followed by a major disjuncture, such as a long pause. It comprises one or more 'intermediate phrases' that are followed by a minor disjuncture, such as a short pause. Then, the intonation contours are described using a sequence of high (H) and low (L) tones, ordered from the beginning of an utterance to the end. Pitch accents (e.g., H*, L*, L+H*, and L*+H) are aligned with a stressed syllable. For bitonal pitch accents, the * indicates the tone that is directly aligned with the stressed syllable. The nuclear accent is the final pitch accent in the intonational phrase and is, in theory, perceived as the most

prominent pitch accent in the phrase. Phrase accents (e.g., H- and L-) are aligned at the right edge of the intermediate phrase to indicate the relative pitch level there. Boundary tones (e.g., H% and L%) are aligned at the right edge of the intonational phrase to indicate the presence of an intonational phrase-level tone[7]. Finally, a tune comprises at least one pitch accent, a phrase accent, and a boundary tone (e.g., H* L- L%).

The annotations that appear in the intonation examples in this thesis used the Tones and Break Indices (ToBI) transcription system, which is based on the autosegmental-metrical theory: MAE_ToBI for English (Beckman, Hirschberg, & Shattuck-Hufnagel, 2005), C_ToBI for Cantonese (Wong, Chan, & Beckman, 2005), and M_ToBI for Mandarin (Peng, Chan, Tseng, Huang, Lee, & Beckman, 2005). MAE_ToBI (abbreviated as ToBI) transcription includes the tones tier and the words tier.[8] The tones tier is used to annotate the pitch accents, phrase accents, and the boundary tones of an intonation contour. The words tier is used to annotate the orthographic spelling of each word in the utterance. Both C_ToBI and M_ToBI differ from ToBI in many of their annotation conventions. First of all, C_ToBI and M_ToBI have a syllables tier and a romanization tier, respectively, for annotating the romanization of the Chinese characters. Secondly, they do not have pitch accents but mark the lexical tones of the syllables in the tones tier. Moreover, M_ToBI has additional symbols for marking pitch range effects in the tones tier (e.g., %q-raise, %-reset, and %e-prom) and C_ToBI also has other boundary tones (e.g., % to indicate the absence of a boundary tone).

[7] Other pitch accents and boundary tones have been defined for English and other languages, but they are beyond the scope of this thesis. See Jun (2005) and Jun (2014).
[8] MAE_ToBI, C_ToBI, and M_ToBI transcriptions also have a break-indices tier. Since break indices have little relevance in this thesis, they have been omitted in the transcriptions in this thesis.

2.6.2 English, Cantonese, and Mandarin Intonation Patterns

English, Cantonese, and Mandarin signal statements and echo questions using distinct intonation patterns. However, they share a general pattern among many other languages: lower pitch (in some form) in statements and higher pitch (in some form) in questions (Bolinger, 1964, 1979; Gussenhoven & Chen, 2000; Ladd, 2008). Nevertheless, the timing, duration, slope, and pitch height of the final fall or rise in statements and questions differ among these languages.

Typically, English statements end with a fall in pitch, whereas echo questions end with a rise in pitch (Wells, 2006: 45). For example, Figure 2.6 shows a paired statement and echo question produced by a native English speaker in my study: "Ann is teacher".

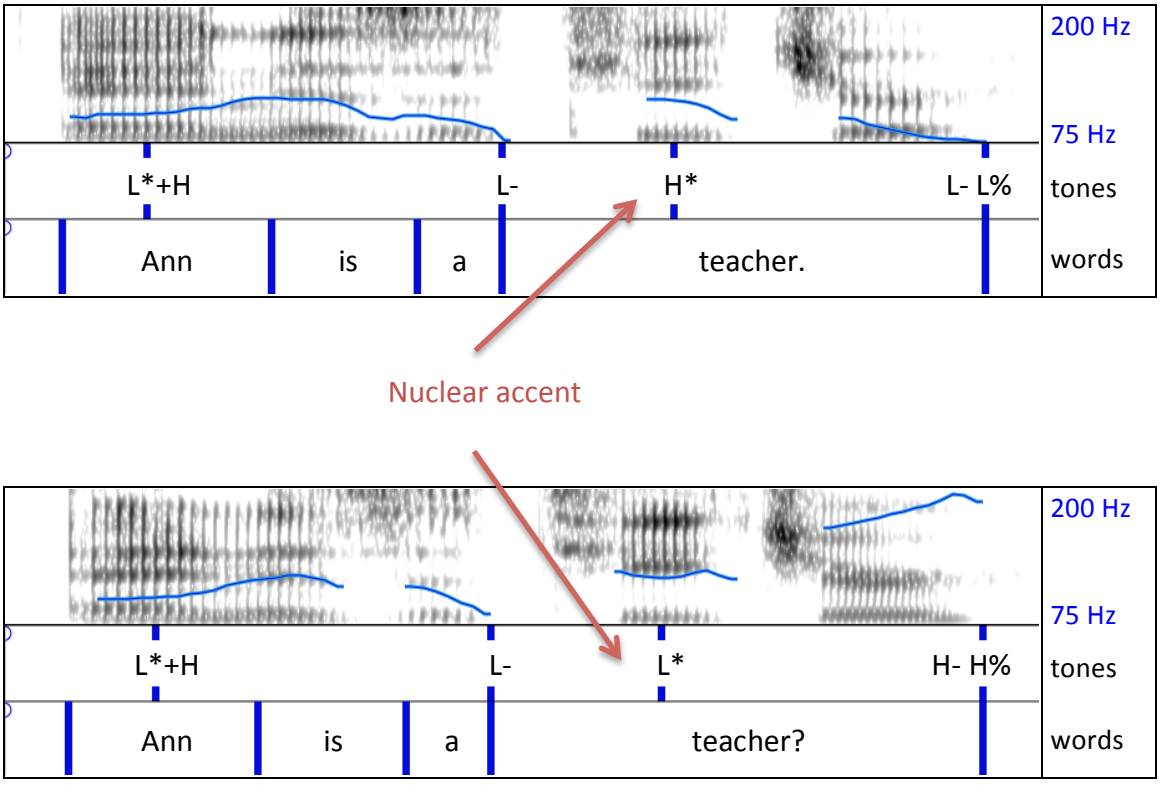


Figure 2.6. Intonation patterns of an English statement and echo question produced by a male speaker.

Both utterances consist of two intermediate phrases: “Ann is a” and “teacher”. In the second intermediate phrase, the tune H* L- L%, which marks a final fall in intonation, denotes a statement in English (top), and the tune L* H- H%, which marks a final rise in intonation, denotes an echo question in English (bottom) (Beckman & Hirschberg, 1999; Pierrehumbert & Hirschberg, 1990). The nuclear accents are H* and L*, approximately where the final fall and rise begin, respectively. Although H* marks a ‘high’ pitch accent, it does not necessarily have a higher F0 value than L*, a ‘low’ pitch accent, as these examples show, due to factors including variation in pitch range between productions and sentence types. Also, the timing of the final fall or rise can differ between utterances due to variation in talker speed, sentence length, and the stress pattern of the final phrase.

In addition, community and individual variation exists in intonation patterns. Some North American speakers (Ladd, 2008) and New Zealanders (Warren, 2005) produce statements with a rising intonation, in a phenomenon commonly known as ‘uptalk’. This high rising terminal creates potential confusion in discriminating between declarative statements and questions. However, Di Gioacchino and Jessop (2011) found that the pitch height of an uptalk rise is not as steep as the pitch height of a question rise.

Similar to English, Cantonese echo questions end with a high F0 rise, regardless of the tone on the final syllable (Flynn, 2003; Fok-Chan, 1974; Gu, Hirose, & Fujisaki, 2005). They show no F0 global raising effect (Xu & Mok, 2011). Cantonese statements, however, retain the pitch direction of the tone on the final syllable. For example, Figure 2.7 shows a paired statement and echo question produced by a native Cantonese speaker in my study: *Wong1 Ji6 m4 zeon2 si4* ‘Wong Ji is not on time’. Both utterances end with the monosyllabic word *si4* ‘time’, which carries a falling tone (Tone 4). In the statement

utterance, there is no boundary tone attached to the end of the final tone, as indicated by %; the canonical contour of this falling tone remains unchanged. In the question utterance, there is a high boundary tone attached to the end of the final tone, as indicated by H% (Wong et al., 2005); the final falling tone rises at the tail end, appearing as a high rising tone. The final syllable also lengthens due to this additional boundary tone.

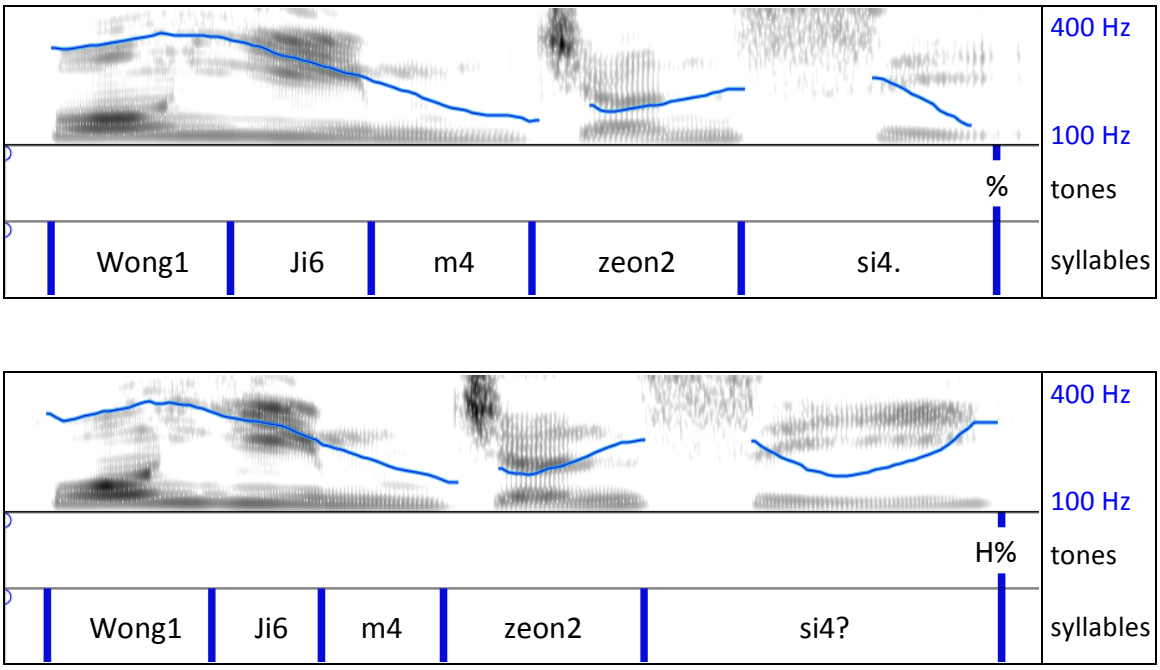


Figure 2.7. Intonation contours of a Cantonese statement and echo question produced by a female speaker: 'Wong Ji is not on time'.

In both statements and echo questions, there is evidence of declination (in pitch) prior to the final syllable. For example, Figure 2.8 shows a paired statement and echo question produced by the same speaker as in Figure 2.7: *Wong1 Ji6 gaau3 lik6 si2* 'Wong Ji teaches history'. In both utterances, the second Tone-6 syllable, *lik6*, is relatively lower in pitch than the first Tone-6 syllable, *Ji6*. Vance (1976) has also found declination in both Cantonese statements and questions in his study of Cantonese tones and intonation. Declination is generally known to occur in statements and partially accounts for their

falling pitch contour in many languages. Declination in questions, however, is less common.

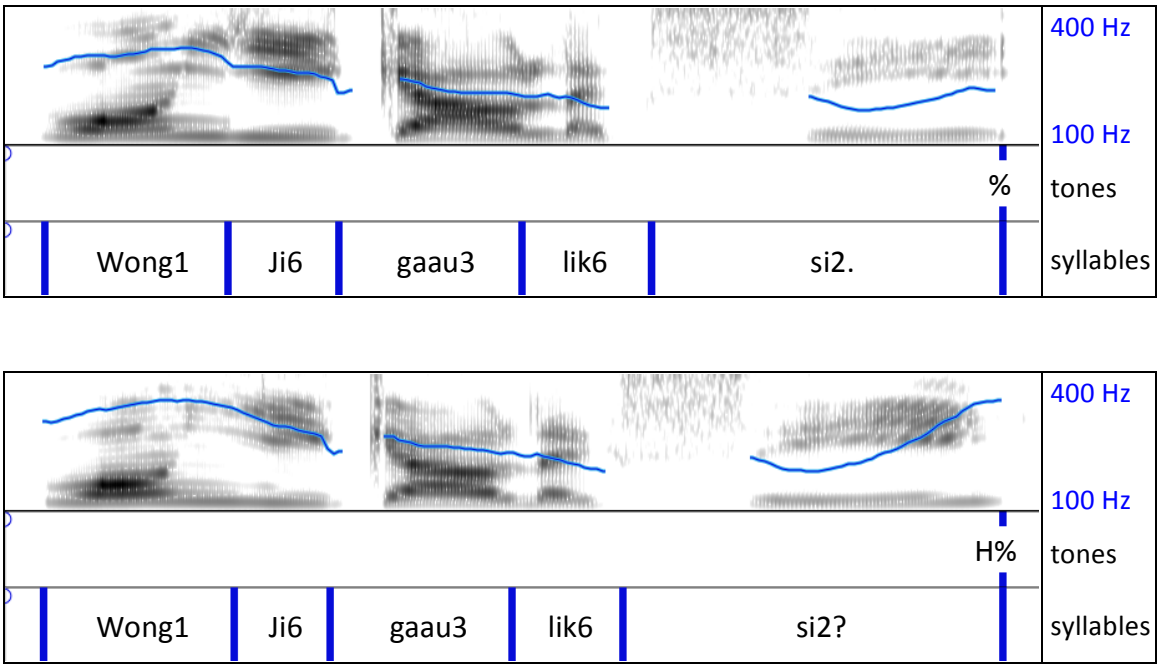


Figure 2.8. Intonation contours of a Cantonese statement and echo question produced by a female speaker: 'Wong Ji teaches history'.

Since statements retain their tonal contours utterance-finally, confusion can potentially arise when listeners discriminate between statements and echo questions in Cantonese. Ma, Ciocca, and Whitehill (2006) investigated the effect of intonation on the perception of lexical tones by native listeners of Hong Kong Cantonese and found that the listeners misperceived many of the tones in the final position of questions as a high rising tone. Their result suggests that listeners may have difficulty disambiguating the relative contributions of tone and intonation on the F0 contour in questions.

Unlike Cantonese, Mandarin statements retain the canonical contour of the final tone. To differentiate between statements and echo questions, native speakers signal Mandarin echo questions with a higher pitch than statements (Peng et al., 2005; Yuan, Shih, & Kochanski, 2002). This global 'raised pitch' effect (Peng et al., 2005) occurs

gradually throughout the utterance. For example, Figure 2.9 shows a paired statement and echo question produced by a native Mandarin speaker in my study: *Wang1 Wu3 kan4 dian4 shi4* 'Wang Wu watches TV'.

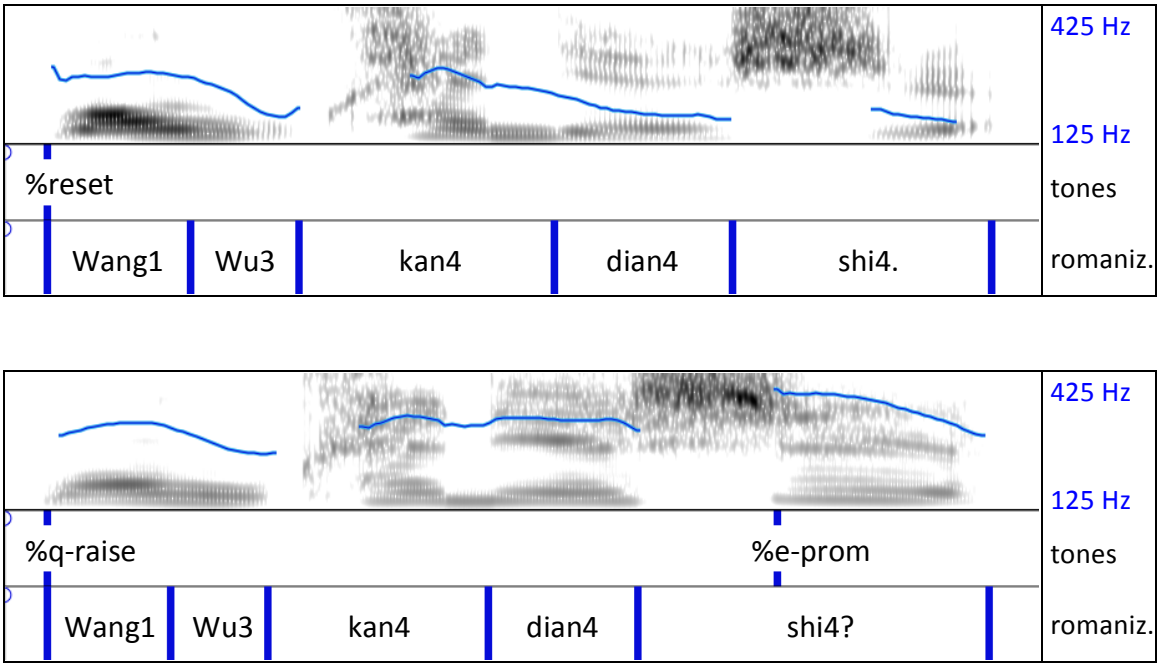


Figure 2.9. Intonation contours of a Mandarin statement and echo question produced by a female speaker: 'Wang Wu watches TV'.

Both utterances end with the syllable *shi4*, which carries a falling tone (Tone 4). In the statement utterance, the pitch range is at the default, neutral level, as indicated by %reset, the pitch range reset symbol. The final tone, Tone 4, is realized as a falling tone. In the question utterance, there is a gradual rise in pitch throughout the utterance compared to the statement. This raised pitch effect is annotated with %q-raise, the question rise marker, at the start of the utterance. The elevated pitch is much higher on the final syllable, accompanied by a local expansion of the pitch range. The pitch range expansion is marked with %e-prom, indicating a local prominence. Similar to Cantonese, the final syllable of the question is lengthened. Declination is also evident in the statement (Shih,

2000; Xu & Wang, 1997), as the sequence of the final three falling tones shows a continually decreasing F0 (i.e., *kan4 dian4 shi4* 'watches TV').

2.7 Summary

This chapter has provided background on the target languages (English, Cantonese, and Mandarin) and the intonation patterns that will be examined in this thesis. The next chapter introduces an exemplar-based model that will be tested on the classification of these fundamental intonation patterns in these three languages.

Chapter 3: Proposed Exemplar-based Model of Intonation Perception

3.1 Overview of the Model

This chapter describes a proposed computational model for simulating an exemplar-based process of categorizing statements and echo questions based on the intonation of naturally produced utterances. It uses a simplified version of the algorithm from Johnson (1997) and Nosofsky (1988). This 'kernel' model, as I call it, assumes that the listener stores all acoustic details of an experienced utterance—including intonation—in 'episodic' memory (a memory system for experienced events) (Tulving, 1972). Since my research questions pertained to statement and echo question intonation only, the model was tested solely on these intonation patterns.

As described in Chapter 2, the statement and echo question intonation patterns of English, Cantonese, and Mandarin are characterized primarily by F0 height, direction, and slope, and the timing of F0 rises and falls. Therefore, the naturally produced utterances that were presented to the model were represented with only their F0 values. By focusing on this dimension alone, the simulation results can reveal cross-linguistic differences that emerge from using F0 to identify statements and echo questions.

This model required training on (or experience with) statements and echo questions before testing, as will be described in Chapter 6. In the spirit of exemplar theory, the F0 contours of the training and test utterances were not normalized to offset speaker variability in pitch between productions and across speakers. The utterances, however, were pre-segmented into sentences. This approach assumes that sentences—or at least their intonation patterns—are stored individually in memory.

3.2 Exemplar-based Process of Categorization

To simulate the exemplar-based process of categorizing statements and echo questions, the model first stores the training stimuli as exemplars in virtual memory according to their categories: a statement or question. Then the model processes each new token from the test stimuli by comparing it with all of the stored exemplars in each category. Through these comparisons, the model calculates the overall similarity value of the incoming token to the exemplars in each category. Based on this overall similarity calculation, the model categorizes the new token as follows. If the overall similarity value for the question category is greater than that for the statement category, the model categorizes the new token as a 'question' and stores it in memory as a question exemplar. Otherwise, the model classifies the new token as a 'statement' and stores it as a statement exemplar. The latter includes the case where the overall similarity values of both categories are equal.[9] The argument for doing so is that listeners tend to default to 'statement' when they cannot distinguish the sentence type of an utterance (Ma, Ciocca, & Whitehill, 2011; Yuan, 2011). Once categorized, the newly stored exemplar is used with the other stored exemplars in the similarity calculations during the processing of subsequent test stimuli.

The algorithm detailed in formulas (3.1) to (3.3) is a simplified version of the algorithm proposed by Johnson (1997) and Nosofsky (1988). To derive the overall similarity value between a new token i and a category C_k (statement or question), first the auditory distance d_{ij} between a new token i and an exemplar j in C_k is calculated based on

[9] In the case where the overall similarity values of both the 'statement' and 'question' categories are equal for a token, I would flag that token. Later, when I checked the results of the model's categorization of statements and questions (as described in Chapter 6), there was no token that had the same overall similarity value for both categories.

their auditory properties x_i and x_j, using the formula in (3.1). Then the auditory similarity s_{ij} between the token and the exemplar is calculated by applying an exponential function to the auditory distance, using the formula in (3.2). This step is to ensure that auditorily 'close' exemplars have a greater impact than auditorily 'distant' exemplars on the overall similarity calculation. Finally, the overall similarity S_{ki} for C_k is the sum of the auditory similarity values between the new token i and all of the individual exemplars in C_k, as in (3.3).

(3.1) Auditory distance: $d_{ij} = [\ \sum(x_i - x_j)^2\ ]^{1/2}$

(3.2) Auditory similarity: $s_{ij} = e^{-d_{ij}}$

(3.3) Overall similarity: $S_{ki} = \sum s_{ij},\ j \in C_k$

Since F0 is a salient cue in signaling statement and question intonation in English, Cantonese, and Mandarin, the model uses F0 values of the tokens and exemplars as the 'auditory properties' in its similarity calculation. I tried to take a balanced approach in determining how many F0 values to feed the model for each exemplar. On the one hand, there needs to be a sufficient number of F0 values to capture the sentence-level question intonation cue in the utterance; on the other hand, too many F0 values could end up capturing the word- or syllable-level tonal variations in the intonation contour for Cantonese and Mandarin. Therefore, I based the number of F0 values on the average syllable length of the test sentences ((5 + 7 + 9 + 11 + 13) / 5 = 9, as will be described in Chapter 4) plus an initial point and a final point. These F0 values ($F0_1$ to $F0_{11}$) are extracted at eleven equidistant time points of the utterance, as shown in Figure 3.1.

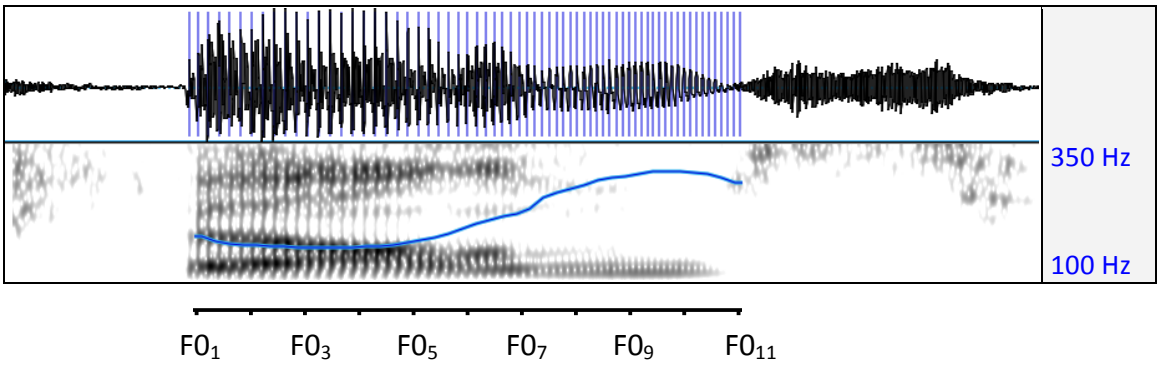


Figure 3.1. F0 values at eleven equidistant time points of an utterance of 'films'. The first point begins at the first voiced cycle of the utterance and the last point ends at the last voiced cycle of the utterance. The points in between are at every 10% of the voiced portion of the utterance. Using these eleven F0 values as auditory properties, the auditory distance d_{ij} is the Euclidean distance between corresponding F0 values of the new token *i* and the stored exemplar *j*, as in (3.4).

(3.4)

$$d_{ij} = \left[\sum_{n=1}^{p}\left(\mathrm{F0}_{ni} - \mathrm{F0}_{nj}\right)^2\right]^{1/2}, \text{ where } p = 11$$

In order to determine how well the kernel model performs across all three languages based on F0 alone (i.e., to determine the saliency of the F0 cue across these languages), the model uses no other acoustic information in the similarity calculation. The simulation results of this model can serve as a benchmark for comparing the results of future simulations that include additional acoustic information in the similarity calculation.

3.3 Auditory Properties for Similarity Calculation

The auditory properties, or the eleven equidistant time points $F0_1$ to $F0_{11}$, were extracted from continuous speech recordings by using Praat (Boersma & Weenink, 2013) as

follows. First, I marked the syllable and sentence boundaries of the target sentences in Praat textgrids for each of the recorded sound files, as shown in Figure 3.2. Then, a Praat script extracted these sentences from the original sound files into individual sentence files. Another Praat script used the 'Pitch' and 'PointProcess' objects in Praat to determine the beginning and end of the periodicity of the utterance in each sentence file: the 'Pitch' object extracted the F0 contour from the sound file, and the 'PointProcess' object converted the 'Pitch' object into a sequence of glottal pulses corresponding to the timing and frequency of the F0 contour. The first and last points of the 'PointProcess' object provided the actual times (in seconds) for calculating the locations of the eleven equidistant points in the F0 contour of the utterance, as shown in Figure 3.3.

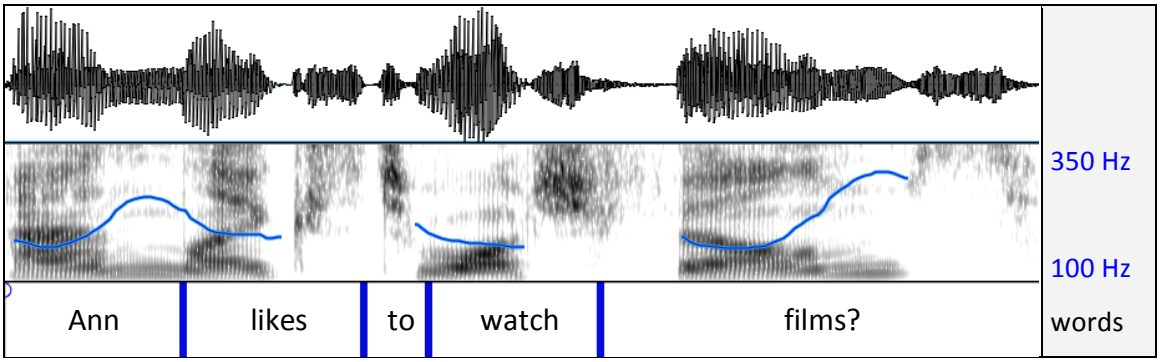


Figure 3.2. Annotation of an English question produced by a female speaker.

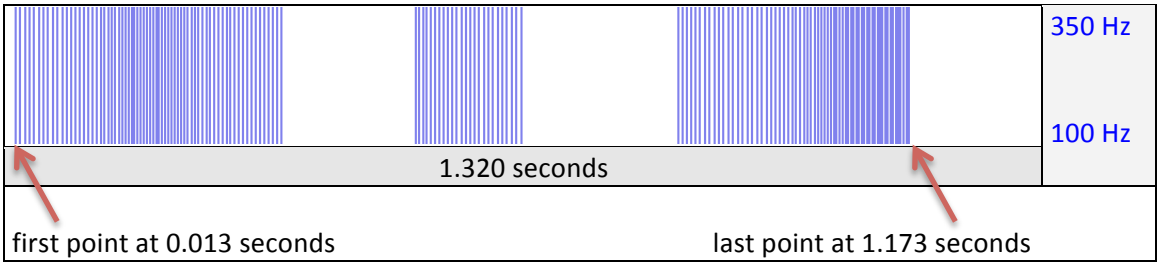


Figure 3.3. 'PointProcess' object in Praat for the utterance in Figure 3.2.

Since part of an utterance can be voiceless or aperiodic—as the example in Figure 3.3 shows, there can be discontinuities in the F0 contour for the entire utterance. To approximate the F0 contour of a discontinuous portion of an utterance, linear interpolation (Praat's 'interpolate' function) was applied to the 'Pitch' object, as shown in Figure 3.4.

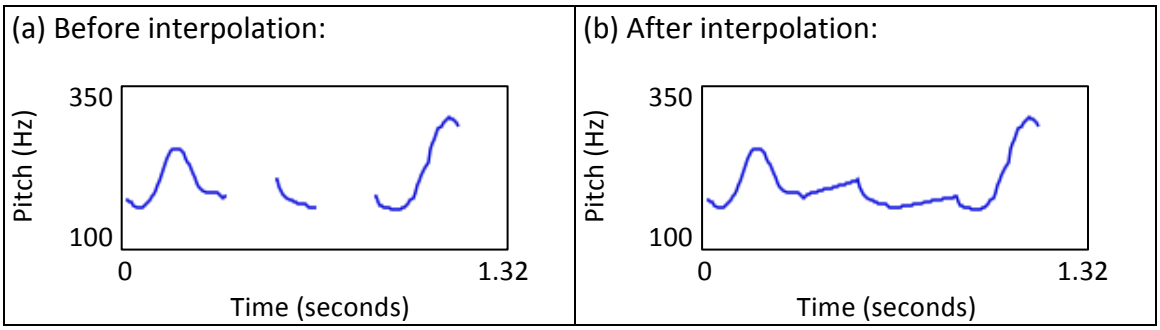


Figure 3.4. F0 contour of the utterance in Figure 3.2 before and after applying interpolation.

Finally, the F0 values of the eleven equidistant time points defined in (3.5) were extracted from the interpolated F0 contour of the utterance, as shown in Figure 3.5, to be used as auditory properties.

(3.5) Time point: $t_i = t_1 + (i - 1) * (t_{11} - t_1)$ where $i \in [1..11]$ and t_i = the time value of point i

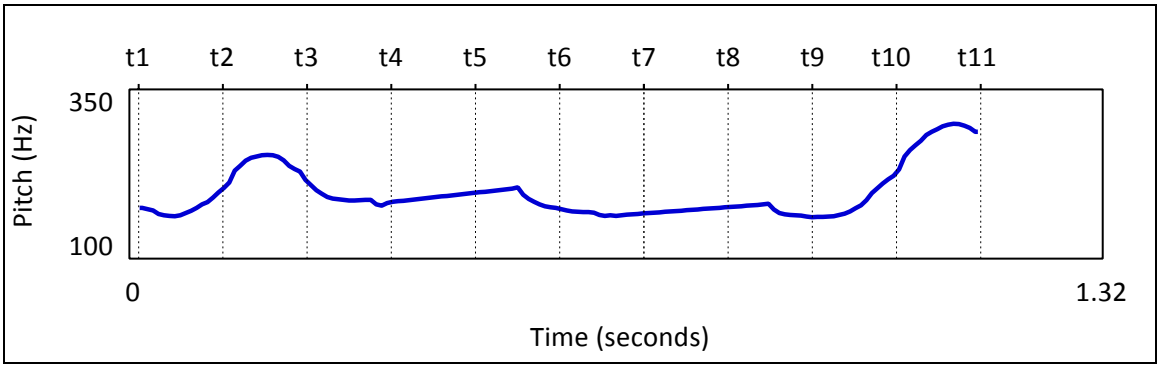


Figure 3.5. Eleven equidistant time points of the pitch contour in Figure 3.4(b).

3.4 Training and Testing Using Cross-Validation

As mentioned earlier, the purpose of the model was to simulate an exemplar-based process of intonation perception to test one account of how listeners perceive variation in speech. To that end, the model was tested in the same way as the human listeners in the perception study described in Chapter 5. Since these human listeners were fluent native speakers of the various languages being tested, they had prior experience with their test language. To parallel this reality, the model would require prior experience with stimuli/sentences from each test language before categorizing new tokens. This experience was thus simulated during the model's training process. How much training to give to the model depended on the method used to validate the testing.

In a scenario where there were 100 tokens, if the model was trained on 99 tokens and then tested on one token, it would likely perform better than if it was trained on one token and then tested on 99 tokens. The model in the first case benefits from having much more experience or exemplars in memory to compare with the new token than the model in the second case. However, in the first case, the model could still perform poorly if the new token was atypical in comparison with the 99 exemplars in memory. Thus, there is a potential downside in training the model to categorize tokens perfectly: the knowledge that the model gains in training might be too specific and only capable of categorizing the tokens presented in training, without being able to recognize the more general structures of the categories. For example, if the model never experienced statements that end in a rising tone [25] when trained to categorize statements and echo questions in Cantonese, it would assume that all sentences with a high final F0 rise in Cantonese were questions. This problem is known as 'overfitting'.

To avoid overfitting the model, I used k-fold cross-validation (Refaeilzadeh, Tang, & Liu, 2009) to train the model and to evaluate how well it generalizes. K refers to the number of folds used. Figure 3.6 displays a three-fold method to show how k-fold cross-validation works in general.

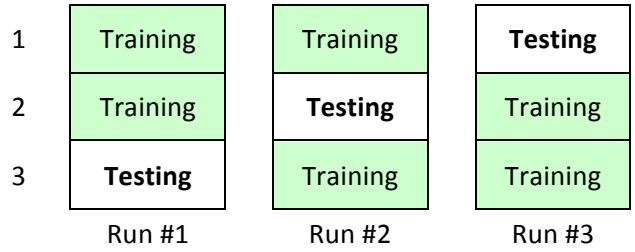


Figure 3.6. Three-fold cross-validation.

In a three-fold cross-validation, the data are divided up into three equal portions. There are three training and test runs. In each run, two-thirds of the data are used for training while the remaining third is used for testing. The training and test data are separate in a given run but cross-over in successive runs so that each token is eventually presented once (and only once) to the model in testing. One way for the model to acquire experience of, or to be trained on, statements and questions is by choosing one token in the data set to be a new token and treating the remaining tokens in the set as exemplars in memory (i.e., the 'leave-one-out cross-validation'). This approach uses k-fold cross-validation where k equals the number of tokens in the data set. In general, training and performance are directly related to one another: more training tokens result in better performance but have a higher risk of overfitting. To maximize both language experience and the amount of test data for the model, a two-fold cross-validation was used in training and testing the model in Chapter 6.

3.5 Summary

This chapter has presented a basic exemplar-based computational model that was used to test whether exemplar theory could account for the perception of different sentence types, based on the intonation or F0 contours of naturally produced utterances. The next chapter describes a production study, in which statements and echo questions produced by native speakers were recorded for use as stimuli for both the perception study (Chapter 5) and simulations of the model (Chapter 6).

Chapter 4: Production Study

4.1 Goals

The goals of this production study were 1) to provide stimuli for testing the computational model and the human listeners in the perception experiment, 2) to generate acoustic measurements to help interpret the perceptual results, and 3) to develop general information on language-specific and cross-linguistic intonation patterns in three languages to help refine the model.

4.2 Methods

4.2.1 Participants

Forty-two native speakers (aged 18-35) participated in the production study: 16 English speakers, 10 Cantonese speakers[10], and 16 Mandarin speakers. The native English speakers had lived in Canada all their lives, except for two speakers who moved to Canada at the age of six to seven months. The native Cantonese speakers were born and raised in Hong Kong, except for one speaker who moved to Hong Kong at the age of two. The native Mandarin speakers were from different regions of China, excluding Hong Kong. Table 4.1 shows the demographic details for these speakers.

Table 4.1. Demographics of the speakers in the production study.

Language	Number of		Age (years)		Age Range (years)		
	Male	Female	Mean	SD	18-23	24-29	30-35
English	8	8	19.31	1.62	16	0	0
Cantonese	5	5	23.00	1.49	7	3	0
Mandarin	8	8	24.94	4.80	7	5	4

[10] It was difficult to recruit native Cantonese speakers from Hong Kong in Calgary, Canada when I ran this production study in Fall 2014. Hence, the numbers of participants were imbalanced across languages.

Nine other speakers also participated in this study, but their recordings were not used for the following reasons. One Cantonese speaker had used a hearing aid at a young age. Between the two excluded Mandarin speakers, one was over the age of 35, and the other was from Taiwan. Among the six excluded English speakers, one was a non-native speaker, one was over the age of 35, one spoke in a child-directed manner, and the remaining three speakers had lived outside of Canada for more than three years.

The participants were recruited from the Introduction to Linguistics course or from flyers posted at the University of Calgary. They were fluent in speaking and reading their native languages, and reported no visual, speech, or hearing impairments. The Cantonese and Mandarin participants were also fluent in reading and understanding Chinese characters in addition to English. For their participation, each speaker received either 1% course credit or $15.

4.2.2 Stimuli

The stimuli were designed to provide variability for testing the exemplar-based model. They comprised five blocks of four dialogues. Each dialogue included a target pair of sentences: a statement and an echo question[11] that were lexically and syntactically identical, as the example in (4.1) shows.

(4.1)	a.	Ann is a teacher.	(statement)
	b.	Ann is a teacher?	(echo question)

Using identical forms for the pair avoided lexical or syntactic cues that listeners could use to identify the different sentence types. To provide the speakers with a dialogue context in which to produce the target sentences, a filler question preceded the target statement,

[11] Unless stated otherwise, 'question' alone refers to 'echo question' from here on.

while a filler affirmative statement followed the target echo question, as the dialogue in (4.2) shows.

(4.2)	a.	Who is Ann?	(filler question)
	b.	Ann is a teacher.	(target statement)
	c.	Ann is a teacher?	(target echo question)
	d.	Yes, Ann is a teacher.	(filler affirmative statement)

The filler questions for the four dialogues in each block were three *wh*-questions (*who*, *what*, and *why*) followed by a yes/no question, as the examples in (4.3) show. The yes/no question was syntactically marked with subject-auxiliary inversion in English, the '*ma*' question marker in Mandarin, and the '*maa*' question marker in Cantonese. The filler questions were varied in order to spread the speakers' attention across all questions and thus prevent them from overemphasizing the echo questions in their readings.

(4.3)	a.	Who is Ann?	(dialogue 1)
	b.	What does Ann teach?	(dialogue 2)
	c.	Why isn't Ann here?	(dialogue 3)
	d.	Does Ann like to watch films?	(dialogue 4)

Shih (2000) found that initial pitch is higher on longer sentences than on shorter sentences in Mandarin. This pitch variation is also evident in other languages, including Swedish (Bruce, 1982) and Dutch (Van Heuven, 2004). To offset the effect of sentence length on intonation, the target sentences in blocks A, B, C, D, and E were 5, 7, 9, 11, and 13 syllables long, respectively, as the examples in (4.4) show.

(4.4)	a.	Ann is a teacher.	(block A, 5 syllables)
	b.	Mary is a good dentist.	(block B, 7 syllables)
	c.	Alice is an old high school friend's Mom.	(block C, 9 syllables)
	d.	Andrew is an electrical engineer.	(block D, 11 syllables)
	e.	Morris is a member of the English Student Club.	(block E, 13 syllables)

As described in Chapter 2, the English question rise is generally aligned with the nuclear accent (or the final stressed syllable) of the intonational phrase. To test the effect

of the timing of the question rise on the model, the English target sentences ended in words that varied in syllable length and stress pattern. That is, half of the sentences ended in a monosyllabic word, and among the remaining half of the sentences, some of the final polysyllabic words were stressed on the final syllable. In total, 65% of the English sentences ended with a stressed syllable, as in (4.5a), and 35% ended with an unstressed syllable, as in (4.5b).

(4.5) a. Ann likes to watch **films**? [ˈ**fɪlmz**]
b. Ann teaches **history**? [ˈhɪs.**t͡ʃɹi**]

As for the two tone languages, to reduce segmental effects of the sentence-final syllable on intonation, the target pairs of each block ended with a different syllable: *shi*, *yi*, *ma*, *fu*, and *fen* for Mandarin, and *si*, *ji*, *maa*, *fu(k)*, and *fan* for Cantonese. In addition, to balance the effects of different lexical tones on intonation for Mandarin and Cantonese, the four target pairs within each block ended in a different tone, as the Mandarin target statements in (4.6) show.

(4.6) a. *Wang1 Wu3 shi4 lao3.**shi1***. (dialogue 1 – ending in *shi*, Tone 1)
Wang Wu is teacher
'Wang Wu is a teacher.'

b. *Wang1 Wu3 jiao4 li4.**shi3***. (dialogue 2 – ending in *shi*, Tone 3)
Wang Wu teach history
'Wang Wu teaches history.'

c. *Wang1 Wu3 bu4 zhun3 **shi2***. (dialogue 3 – ending in *shi*, Tone 2)
Wang Wu not accurate time
'Wang Wu is not on time.'

d. *Wang1 Wu3 kan4 dian4.**shi4***. (dialogue 4 – ending in *shi*, Tone 4)
Wang Wu watch TV
'Wang Wu watches TV.'

Cantonese has six contrastive tones, two more than Mandarin. To be able to better compare the model's performance on the final lexical tones between these two languages,

the six Cantonese tones were first combined into four groups and then these groups were mapped onto the four Mandarin tones. Table 4.2 shows the mapping scheme. Cantonese's Tone 2 and Tone 5 were grouped together as an R tone, while Tone 3 and Tone 6 were grouped together as an L tone. The rationale for grouping these tones in this way is that it is difficult, even for native speakers, to produce the tones in each pair distinctively (Bauer & Benedict, 1997) and to perceive their differences (Ciocca & Lui, 2003). Mok, Zuo, and Wong (2013) have found that these tones are merging among young speakers between the ages of 18 and 22.

Table 4.2. Mapping between Mandarin and Cantonese tones.

Tone Letter	Shape	Mandarin	Cantonese
H	High	Tone 1 [55]	Tone 1 [55]
R	Rising	Tone 2 [35]	Tone 2 [25], Tone 5 [23]
L	Low	Tone 3 [214]	Tone 3 [33], Tone 6 [22]
F	Falling	Tone 4 [51]	Tone 4 [21]

Table 4.3 shows the syllables and tone letters of the sentence-initial and sentence-final syllables of the Mandarin and Cantonese target sentences. The initial two syllables of every sentence were names of people and remained the same throughout each block. For blocks A to D, these names consisted of the highest pitch level of 5 and the lowest pitch level of 1 in Chao's (1947, 1948) tonal system. These pitch extremes displayed the extent of the speaker's pitch range at the beginning of the sentence and could thus serve as a cue for distinguishing questions from statements. For block E, the names carried two high tones, which could also possibly cue listeners to the sentence type. Where possible, the same final syllable was used in corresponding sentences between both languages. Even so, the tone of that target syllable sometimes differed between these languages, for example, 椅 'chair' is *ji2* (R) in Cantonese and *yi3* (L) in Mandarin.

Table 4.3. Mandarin and Cantonese target syllables and tones, initially and finally.

Block	Dialogue	Sentence-initial		Sentence-final	
		Mandarin	Cantonese	Mandarin	Cantonese
A	1	*Wang1 Wu3* (H+L)	*Wong1 Ji6* (H+L)	*shi1* (H)	*si1* (H)
	2			*shi3* (L)	*si2* (R)
	3			*shi2* (R)	*si4* (F)
	4			*shi4* (F)	*si6* (L)
B	1	*Ye4 Shi2* (F+R)	*Jyu4 So2* (F+R)	*yi1* (H)	*ji1* (H)
	2			*yi3* (L)	*ji2* (R)
	3			*yi2* (R)	*ji4* (F)
	4			*yi4* (F)	*ji3* (L)
C	1	*Li3 Yi1* (L+H)	*Lou6 Faa1* (L+H)	*ma1* (H)	*maa1* (H)
	2			*ma3* (L)	*maa5* (R)
	3			*ma2* (R)	*maa4* (F)
	4			*ma4* (F)	*maa6* (L)
D	1	*Wu2 Er4* (R+F)	*Heoi2 Wu4* (R+F)	*fu1* (H)	*fu1* (H)
	2			*fu3* (L)	*fu2* (R)
	3			*fu2* (R)	*fuk6* (L)
	4			*fu4* (F)	*fu4* (F)
E	1	*Su1 San1* (H+H)	*Sou1 Sin1* (H+H)	*fen4* (F)	*fan6* (L)
	2			*fen2* (R)	*fan4* (F)
	3			*fen1* (H)	*fan1* (H)
	4			*fen3* (L)	*fan2* (R)

Finally, the corresponding English, Cantonese, and Mandarin sentences in each block were similar in semantic context or meaning. As much as possible, the stimuli were composed of commonly or frequently used words. The production sentences for all three languages are listed in Appendix A.

4.2.3 Procedure

The participants first completed a brief questionnaire about their language background (see Appendix B) and then performed a reading task. During the reading task, the stimuli were presented to the participants in Microsoft PowerPoint on an iMac computer. Each dialogue was displayed on a separate slide as a conversation between two speakers: A and B. The participants were instructed to play the roles of both speakers, imagining that they were talking to a friend. The participants were also instructed to express the echo question in the dialogue as a confirmation of the previous statement and not as a surprise; surprise increases the emotion of an utterance, which could affect the intonation of the echo question (Paeschke, 2004).

(a)

A:	What does Ann teach?
B:	Ann teaches history.
A:	Ann teaches history?
B:	Yes, Ann teaches history.

(b)

A:	汪義教乜嘢?
B:	汪義教歷史。
A:	汪義教歷史?
B:	係,汪義教歷史。

(c)

A:	汪五教什么?
B:	汪五教历史。
A:	汪五教历史?
B:	是,汪五教历史。

Figure 4.1. English, Cantonese, and Mandarin dialogues presented to the speakers.

The English sentences were displayed in English orthography. The Mandarin sentences were displayed in simplified Chinese characters, while the Cantonese sentences were displayed in traditional Chinese characters. For all three languages, the texts were displayed horizontally from left to right. Figure 4.1 shows a dialogue in English, followed by the corresponding dialogues in Cantonese and Mandarin.

The participants were recorded individually in a sound-attenuated booth in the Phonetics Lab at the University of Calgary. They read aloud into a Shure SM-48 microphone, which was placed approximately four inches from their mouths. A KayPentax CSL 4500 device recorded each reading, converted the analog signal to digital, and then outputted the digital signal to another computer that was running Adobe Audition. Audition captured the digital signal at a sampling rate of 48 kHz in a 16-bit mono channel.

The recording session generally lasted 45 minutes. The participants read two practice dialogues prior to the main dialogues to ensure that they understood the instructions and to check their recording volume. The sentences in the practice dialogues were lexically different from the sentences in the main dialogues. The participants were asked to read the sentences naturally at their normal talking speed and volume. If they accidentally misread a word or phrase, they would re-read the entire dialogue. All five blocks were counterbalanced among the speakers. Each participant read each dialogue three times. The initial reading was treated as a practice run to familiarize the speakers with the target sentences and was discarded. The first and second repetitions were saved

to a sound file in .wav format as readings 1 and 2 for acoustic analysis[12] in Praat (Boersma & Weenink, 2013).

4.2.4 Acoustic Analysis

The acoustic analysis examined the F0 contours of the recorded sentences in order to determine which parts of the sentence might contain the salient F0 cues for differentiating statements and echo questions. Each language was analyzed separately. The mean F0 contours of the statements and echo questions were approximated from the F0 measurements at eleven equidistant time points of the utterance (as described in Chapter 3). Due to devoicing or creaky voice, 115 pairs of utterances (63 from English, 13 from Cantonese, and 39 from Mandarin) had to be excluded from this analysis because Praat was unable to extract their F0 values. The remaining pairs totaled to 725.

4.2.5 Exemplar-based Classification

To determine at which time points of the utterance significant differences between statements and echo questions emerged, I compared the accuracy rates of the proposed exemplar-based model (Chapter 3) in classifying statements and echo questions. I tested the model using two separate classification methods to find out which method yielded better performance. The model was presented with the 725 pairs of statements and echo questions. Each pair, in turn, served as two new tokens for classification while the remaining 724 pairs served as exemplars in memory.

[12] Reading 2 was not used in the analysis presented in this thesis, but part of its data was analyzed and reported in Chow and Winters (2015).

The *Single-point Classification* (SpC) method tested the model in 11 conditions using the F0 value at a single time point in each condition. Specifically, condition 1 calculated the auditory distance d_{ij} between token *i* and exemplar *j* in memory using only $F0_{\mathbf{1}}$, while condition 2 used only $F0_{\mathbf{2}}$, condition 3 used only $F0_{\mathbf{3}}$, and so forth, as shown in (4.7).

(4.7) Single-point Classification Method: F0 value at a single time point

Condition 1: $d_{ij} = [\ (F0_{\mathbf{1}i} - F0_{\mathbf{1}j})^2\ ]^{1/2}$

Condition 2: $d_{ij} = [\ (F0_{\mathbf{2}i} - F0_{\mathbf{2}j})^2]^{1/2}$

Condition 3: $d_{ij} = [\ (F0_{\mathbf{3}i} - F0_{\mathbf{3}j})^2]^{1/2}$

…

Condition 11: $d_{ij} = [\ (F0_{\mathbf{11}i} - F0_{\mathbf{11}j})^2]^{1/2}$

The formula for (4.7) is shown in (4.8).

(4.8) Single-point Classification Method: F0 value at a single time point

Condition *p*:

$$d_{ij} = \left[\left(F0_{pi} - F0_{pj}\right)^2 \right]^{1/2}$$

where $p \in [1..11]$

The *Multi-point Classification* (MpC) method tested the model in 11 conditions using the F0 values at accumulated successive time points. Condition 1 calculated the auditory distance d_{ij} between token *i* and exemplar *j* in memory using $F0_{\mathbf{1}}$ only. Then each successive condition added the next F0 value to the previous condition. That is, condition 2 applied both $F0_{\mathbf{1}}$ and $F0_{\mathbf{2}}$ to the calculation of auditory distance, condition 3 applied $F0_{\mathbf{1}}$, $F0_{\mathbf{2}}$, and $F0_{\mathbf{3}}$ to the calculation, and so forth, as shown in (4.9).

(4.9) Multi-point Classification Method: F0 values at successive time points

Condition 1: $d_{ij} = [\ (F0_{\mathbf{1}i} - F0_{\mathbf{1}j})^2\ ]^{1/2}$

Condition 2: $d_{ij} = [\ (F0_{\mathbf{1}i} - F0_{\mathbf{1}j})^2 + (F0_{\mathbf{2}i} - F0_{\mathbf{2}j})^2]^{1/2}$

Condition 3: $d_{ij} = [\ (F0_{\mathbf{1}i} - F0_{\mathbf{1}j})^2 + (F0_{\mathbf{2}i} - F0_{\mathbf{2}j})^2 + (F0_{\mathbf{3}i} - F0_{\mathbf{3}j})^2]^{1/2}$

…

Condition 11: $d_{ij} = [\ (F0_{\mathbf{1}i} - F0_{\mathbf{1}j})^2 + (F0_{\mathbf{2}i} - F0_{\mathbf{2}j})^2 + (F0_{\mathbf{3}i} - F0_{\mathbf{3}j})^2 + \ldots + (F0_{\mathbf{11}i} - F0_{\mathbf{11}j})^2]^{1/2}$

The formula for (4.9) is shown in (4.10).

(4.10) Multi-point Classification Method: F0 values at successive time points

Condition p:

$$d_{ij} = \left[\sum_{n=1}^{p}\left(F0_{ni} - F0_{nj}\right)^2\right]^{1/2}$$

where $p \in [1..11]$

Multi-point Classification reveals the time point at which significant differences in F0 emerge between statements and echo questions. In general, it includes more information (F0 values) in its similarity calculation than the Single-point Classification method. However, since it uses accumulated F0 values, it cannot show the saliency of the F0 value at each time point. Single-point Classification reveals the degree of F0 difference between statements and echo questions that the model can detect at each time point. In principle, the classification rates at these points can be used to investigate how much attention listeners are paying to each of these points in order to determine how attention weights in an exemplar model should be incorporated and adjusted in simulations (see Chapter 7).

4.3 Results

4.3.1 Acoustic Analysis

Figure 4.2 shows the mean F0 contours of the sentences produced by the English, Cantonese, and Mandarin speakers, split by speaker gender and sentence type. On average, the male speakers produced these sentences with a lower F0 and a narrower F0 range than the female speakers. This F0 difference between genders is consistent with the myoelastic theory described in Chapter 2. Overall, the intonation of the statements and questions produced by both genders show similar patterns: a final rise in questions and a gradual fall in statements. The white markers on the contours in Figure 4.2 indicate roughly where these contours diverge. Due to a wider F0 range for females, the final diverging point between the statement and question contours is earlier on female-produced sentences than on male-produced sentences for English and Mandarin: at time point 8 (female) versus time point 9 (male) for English and at time point 1 (female) versus time point 6 (male) for Mandarin. For Cantonese, the final diverging point between the statement and question contours is at time point 9, similar to English, on all sentences.

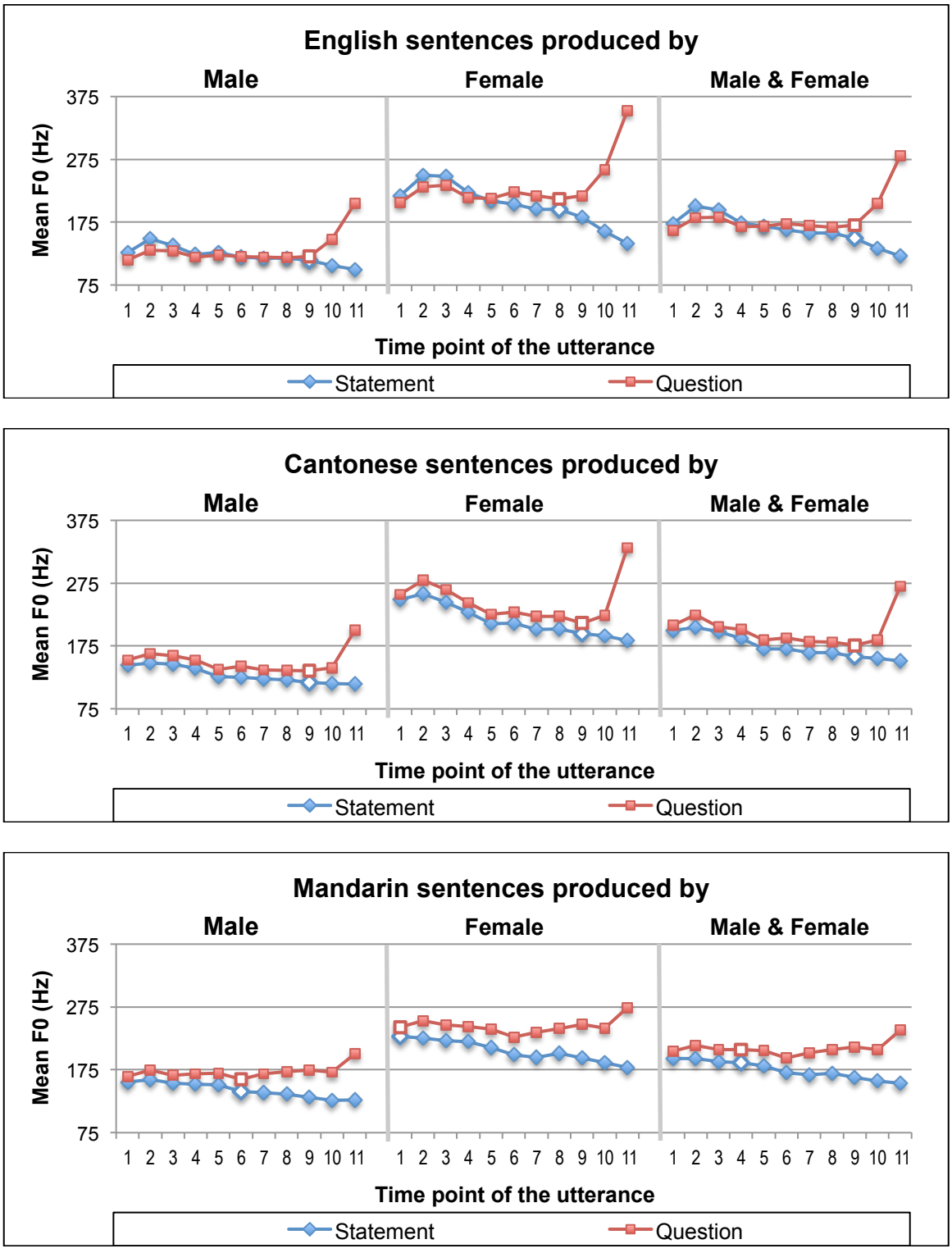


Figure 4.2. Mean F0 contours by speaker gender and sentence type in English, Cantonese, and Mandarin.

Comparing all three languages reveals a general pattern: statements begin at about the same F0 level as questions and then decline below the F0 level of the questions. Interestingly, however, in English, the statement contour rises slightly above the question contour during its initial 30% (between time points 1 and 4) before gradually descending below the question contour. In Cantonese, the statement contour remains slightly below the question contour during the initial 90% of the time (between time points 1 and 10) as the question contour declines along with the statement contour. There is no evidence of elevated pitch in these Cantonese questions, on average. In Mandarin, however, the question contour remains relatively level during the initial 90% of the time (between time points 1 and 10) while the statement contour gradually declines, thus creating an increasingly large gap between these two contours from the start to the end. Finally, although the question contours show a final rise in all three languages, this final rise is much steeper in English and Cantonese than in Mandarin. It is also slightly earlier in English (at time point 9) than in Cantonese and Mandarin (at time point 10). Overall, the F0 patterns produced by the speakers in this study are consistent with what was found in previous literature (see Chapter 2).

The second half of this subsection examines the influence of word stress and lexical tones on statement and question intonation in the three languages. Figure 4.3 shows the mean F0 contours produced by the English speakers, split by final stress and sentence type. On average, the question rise starts earlier on sentences ending in an unstressed syllable (at time point 8) than on sentences ending in a stressed syllable (at time point 9). Since the English question rise starts at or near the nuclear tone, which is anchored to the last stressed syllable of the intonational phrase, the earlier rise on sentences with an unstressed final syllable is because the nuclear tone occurs before the final syllable.

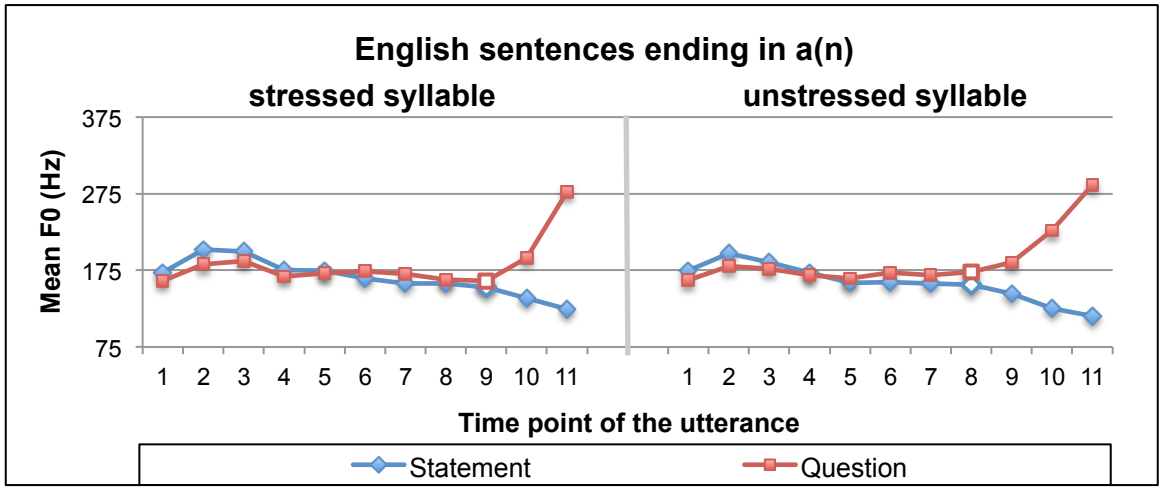


Figure 4.3. Mean F0 contours by final stress and sentence type in English.

Figure 4.4 shows the mean F0 contours produced by the Cantonese speakers, split by final tone and sentence type. Regardless of the lexical tone on the final syllable, all four question contours end with a high F0 rise. The earlier rise on questions ending in a high tone (at time point 9 instead of time point 10) is likely a tonal effect of transitioning from a preceding lower tone to the final high tone. This analysis is based on the fact that the statement contour for sentences ending in a high tone also rises at time point 9 but levels off at time points 10 and 11. The statement contour for sentences ending in a rising tone has a final rise as well, starting at time point 10, but the final height of this rise is relatively low.

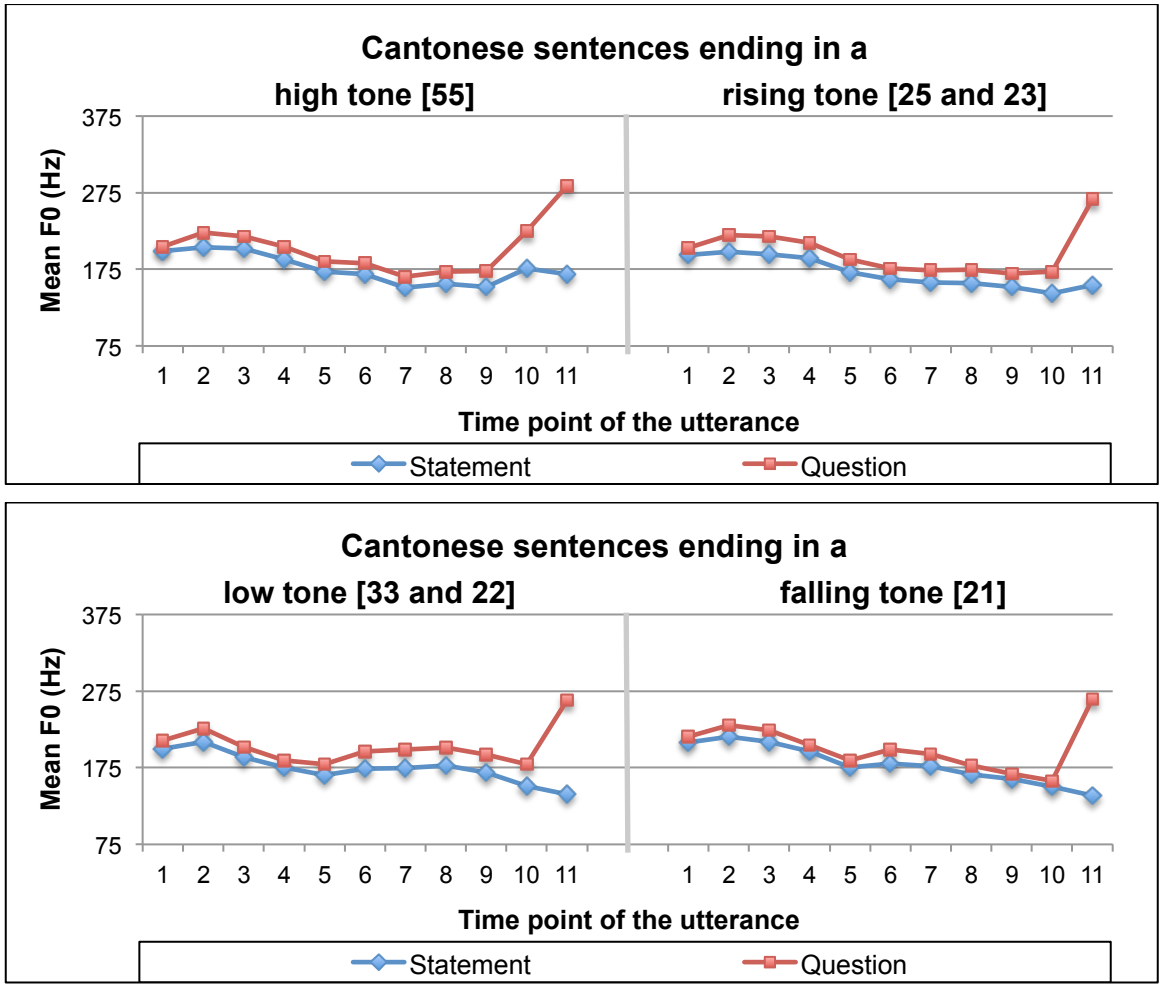


Figure 4.4. Mean F0 contours by final tone and sentence type in Cantonese.

Figure 4.5 shows the mean F0 contours produced by the Mandarin speakers, split by final tone and sentence type. Unlike Cantonese, all four question contours retain the canonical shape of the lexical tone on the final syllable. Here, the final high tones appear to be rising tones due to transitioning from a preceding lower tone, similar to the tonal effect on Cantonese high tones. The final rising tone on the statement contour is relatively low, compared to the rising tone on the question contour. In addition, the final low tone is realized as a [214] tone on questions but as a [21] tone on statements. Finally, the pitch is oddly elevated near the end, at time points 9 and 10, of the question contour ending in a falling tone (see Chow and Winters (2016)).

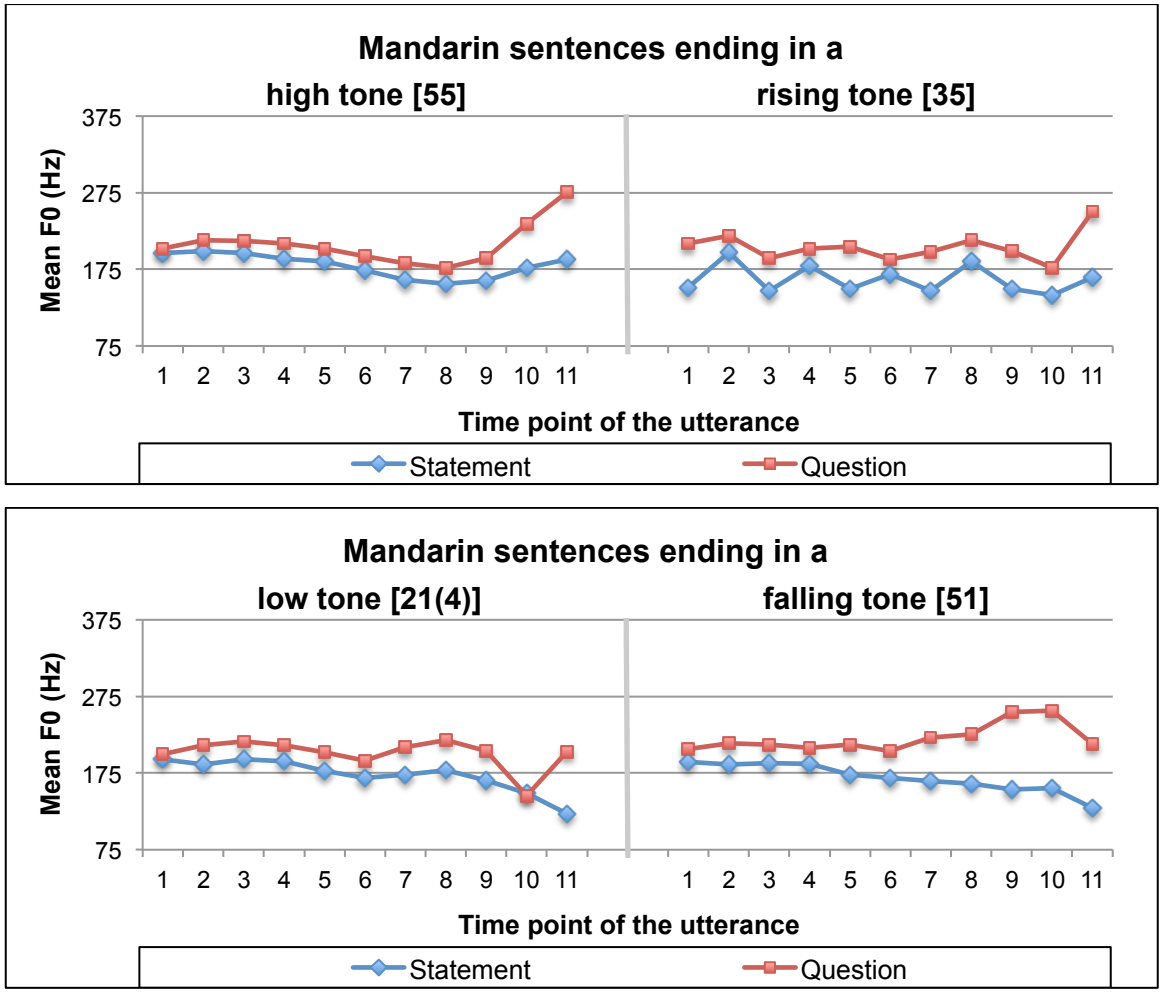


Figure 4.5. Mean F0 contours by final tone and sentence type in Mandarin.

4.3.2 Exemplar-based Classification

Figure 4.6 shows the results of the Single-point Classification and Multi-point Classification of statements and questions in English, Cantonese, and Mandarin. Descriptively speaking, MpC outperformed SpC. Averaged across conditions 1 to 10, SpC identified statements and questions less accurately than MpC (English: SpC = 55%, MpC = 67%; Cantonese: SpC = 56%, MpC = 59%; Mandarin: SpC = 59%, MpC = 61%). On condition 11 alone, SpC also identified statements and questions less accurately than MpC (English: SpC = 82%, MpC = 94%; Cantonese: SpC = 83%, MpC = 95%; Mandarin: SpC = 72%, MpC = 78%). In all six classifications, performance was best in condition 11, which suggests that the most salient cues to the difference between the sentence types reside at the end of the utterance. Based on the best condition (condition 11), classification performance was better on English and Cantonese than on Mandarin, most likely due to the high question rise in English and Cantonese.

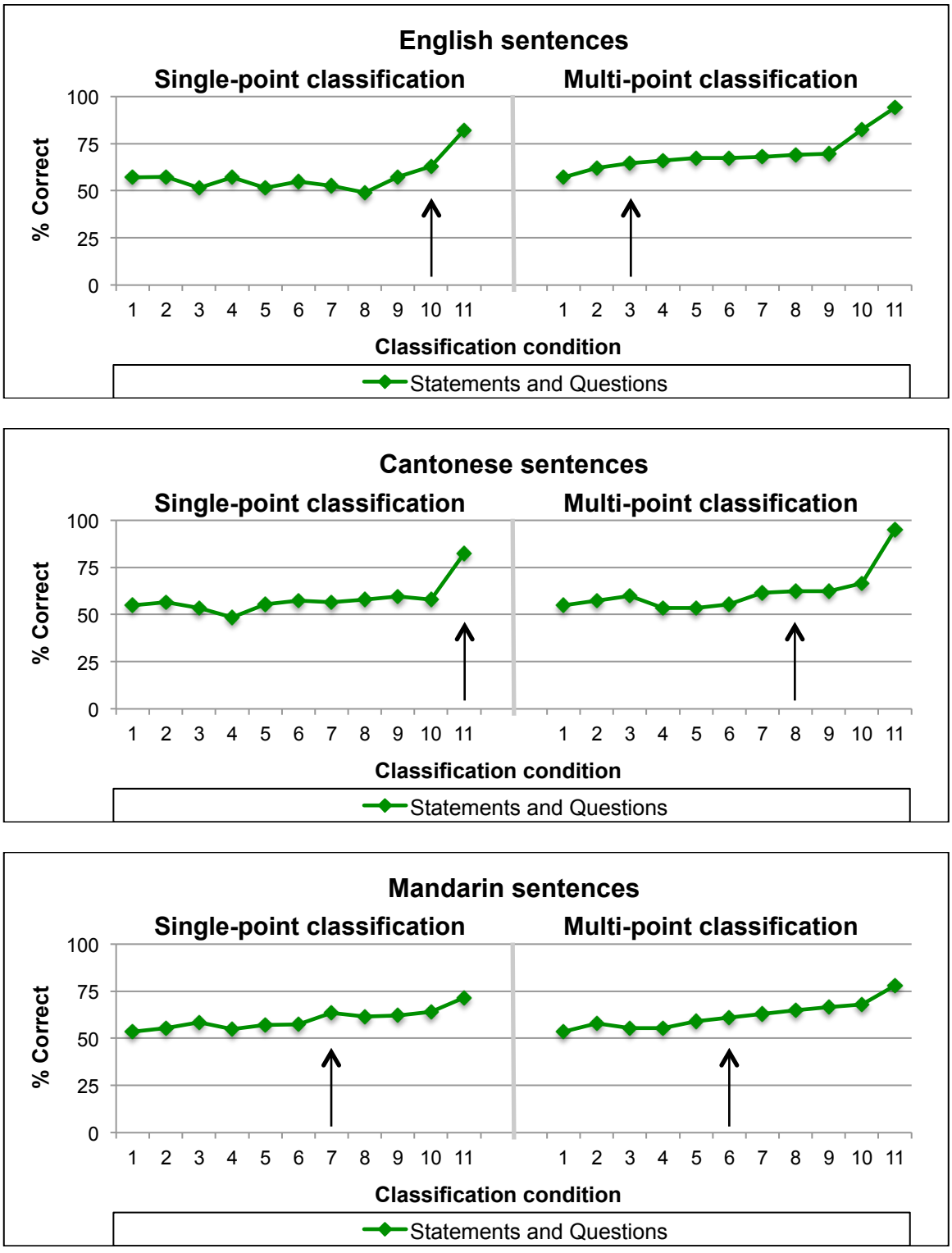


Figure 4.6. Single-point Classification versus Multi-point Classification. Arrows indicate time points where a significant increase in correct classification begins over time point 1.

To determine the effect of time point on the percentage of sentences correctly classified by SpC and MpC, I ran a logistic regression on the percent correct scores from each model. The results of the logistic regression on SpC (Table 4.4) indicate a significant improvement in classification over the initial time point for English [$G^2(10, 5643) = 186.1$, $p < 0.001$] starting at 90% through the utterance in condition 10, for Cantonese [$G^2(10, 4103) = 125.1$, $p < 0.001$] at the last time point of the utterance, and for Mandarin [$G^2(10, 6171) = 63.8$, $p < 0.001$] starting at 60% through the utterance in condition 7. (For English, although condition 8 is significant, the estimate or coefficient is negative, indicating a drop in performance, as shown in Figure 4.6.) These results suggest a correlation between correct classification by SpC and the question intonation contour, specifically the tail-end rises of English and Cantonese and the pitch raising of Mandarin, as shown in Figure 4.2.

Table 4.4. Coefficients and *p*-values of a logistic regression on SpC (* $p < 0.05$).

Single-point Classification	**English**		**Cantonese**		**Mandarin**	
	Estimate	***p*-value**	**Estimate**	***p*-value**	**Estimate**	***p*-value**
Condition 1	0.2661	= 0.003*	0.2039	= 0.050	0.1497	= 0.077
Condition 2	0.0476	= 0.705	0.0433	= 0.769	0.0645	= 0.590
Condition 3	-0.1960	= 0.118	-0.0753	= 0.608	0.1879	= 0.118
Condition 4	0.0079	= 0.950	-0.2681	= 0.068	0.0430	= 0.719
Condition 5	-0.1960	= 0.118	0.0108	= 0.941	0.1296	= 0.280
Condition 6	-0.0631	= 0.615	0.0978	= 0.507	0.1514	= 0.208
Condition 7	-0.1648	= 0.188	0.0433	= 0.769	**0.4049**	< 0.001*
Condition 8	-0.3128	= 0.013*	0.1087	= 0.461	0.3139	= 0.010*
Condition 9	0.0317	= 0.801	0.1748	= 0.237	0.3440	= 0.005*
Condition 10	**0.2592**	= 0.042*	0.1197	= 0.417	0.4358	< 0.001*
Condition 11	1.2703	< 0.001*	**1.3550**	< 0.001*	0.7541	< 0.001*

The results of the logistic regression on MpC (Table 4.5) indicate significant improvement in classification over the initial time point for English [$G^2(10, 5643) = 304.5$, $p < 0.001$] when 20% of the utterance is included in the calculation of auditory distance, for Cantonese [$G^2(10, 4103) = 276.1$, $p < 0.001$] when 70% of the utterance is included, and for Mandarin [$G^2(10, 6171) = 122.8$, $p < 0.001$] when 50% of the utterance is included. These results suggest that, besides the final rise, there are F0 cues earlier in the utterance that can help to distinguish questions for English and Cantonese.

Table 4.5. Coefficients and *p*-values of a logistic regression on MpC (* $p < 0.05$).

Multi-point Classification	**English**		**Cantonese**		**Mandarin**	
	Estimate	***p*-value**	**Estimate**	***p*-value**	**Estimate**	***p*-value**
Condition 1	0.2661	= 0.003*	0.2039	= 0.050	0.1497	= 0.077
Condition 2	0.2095	= 0.099	0.0978	= 0.507	0.1587	= 0.187
Condition 3	**0.3435**	= 0.007*	0.2082	= 0.160	0.0790	= 0.510
Condition 4	0.3864	= 0.003*	-0.0754	= 0.608	0.0718	= 0.549
Condition 5	0.4652	< 0.001*	-0.0754	= 0.608	0.2321	= 0.054
Condition 6	0.4563	< 0.001*	0.0108	= 0.941	**0.3139**	= 0.010*
Condition 7	0.4830	< 0.001*	0.2531	= 0.089	0.3668	= 0.003*
Condition 8	0.5462	< 0.001*	**0.3212**	= 0.031*	0.4669	< 0.001*
Condition 9	0.5645	< 0.001*	0.2984	= 0.045*	0.5301	< 0.001*
Condition 10	1.2571	< 0.001*	0.4732	= 0.002*	0.6109	< 0.001*
Condition 11	2.5507	< 0.001*	2.7807	< 0.001*	1.1122	< 0.001*

4.4 Discussion

Based on the design of the model, I hypothesized that the categorization of statements and questions would rely largely on the width and the length of the F0 gap between statements and questions. Therefore, identifying statements and questions in all three languages would be easiest at the end of the utterance, where the F0 difference between statements and questions is the greatest due to the question rise (i.e., L* L- H% in

English, H% in Cantonese, and %q-raise in Mandarin). Cross-linguistically, identifying statements and questions from the earlier part of utterance should be easier in English and Mandarin than in Cantonese. In English, the nuclear tone can occur before the final syllable, so the question rise and the F0 gap can start in the earlier part of the utterance. For Cantonese, both statements and questions start at nearly the same F0 level, and then the question pattern tends to decline parallel to the statement declination, so the F0 gap between both sentence types is minimal prior to the final question rise. For Mandarin, it should be easier to distinguish both sentence types from the second half of the utterance than from the first half of the utterance, due to the relatively larger F0 gap between statements and questions in the second half. Apparently, the effect of tones on intonation is greater than the effect of stress on intonation, as observed in the greater variation between time points 1 and 8 among the four F0 contours in Figure 4.4 for Cantonese and Figure 4.5 for Mandarin, as opposed to the smaller variation in Figure 4.3 for English. As mentioned in Section 2.2, the acoustic correlate for tone is mainly F0 (Fok-Chan, 1974), while the acoustic correlates for English include F0, duration, and intensity (Adams & Munro, 1978). It is predicted that this tonal effect would decrease the performance of the model in identifying the sentence types in Mandarin as the question cue in Mandarin extends throughout the utterance; this question cue would be affected by a different tone on each syllable throughout the utterance.

Both the observations of the F0 contours and the results of the logistic regression analysis suggest that the F0 gap at the earlier part of the utterance, as well as at the end of the utterance, contributes to the identification of statements and questions in English, Cantonese, and Mandarin. Therefore, the proposed model needs to include this F0

information in its similarity calculation. Only condition 11 of the Multi-point Classification includes the final time point as well as earlier time points of the utterance. In addition, Single-point Classification uses only the F0 value of a single time point in its calculation of auditory distance between a new token and the exemplars in memory, so in essence, it considers only the F0 height and not the F0 contour of the utterance. Multi-point Classification applies F0s at successive time points in its calculation, so it considers F0 height, direction, and range. Therefore, I used condition 11 of the Multi-point Classification in simulating the categorization of statements and questions in Chapter 6. This condition takes into account the most complete representation of the F0 contour and yields the best performance among all 22 conditions, which is evidence that the most complete representation is necessary for better performance. This evidence is in line with exemplar theory, in principle.

4.5 Summary

This chapter has described the production sentences that were used to generate stimuli for the perception study and the simulations of the model. An observational analysis of the F0 contours of the produced sentences showed that the F0 gap between the statement and question contours towards the end of the utterance could be a salient cue in identifying statements and questions in all three languages. A pilot test of the model using two classification methods revealed potential F0 cues in the earlier part of the utterance as well. The next chapter reports on the perception study involving human listeners performing a similar classification task.

Chapter 5: Perception Study

5.1 Goals

The goals of the perception study were 1) to compare how well native listeners can correctly identify statements and questions in English, Cantonese, and Mandarin, based on intonation alone, 2) to find out which parts of an utterance are perceptually salient for each language, 3) to determine the effects of stress and tone on listeners' perception of native sentence intonation, and 4) to provide human perceptual responses to compare with the model's simulation results (described in Chapter 6).

5.2 Methods

This study comprised two sessions. In each session, listeners of each language performed an identification task in which they listened to utterances and indicated whether the token that they had heard was a statement or question. This identification task was implemented as a modified and speeded gating task (Trimble, 2013). The primary goal of the gating task was to identify how much intonational information listeners can get from each gated portion of an utterance. Another goal of the gating task was to bring performance on the identification task down from potential ceiling levels, since native listeners would likely score near 100% if they heard whole sentences. By examining listener performance in less than ideal conditions, this approach makes it more likely for any potential performance differences to emerge between listeners of different languages, or between the model and the human listeners. The task was also designed to potentially identify language-specific differences in intonation perception, such as whether English and

Cantonese listeners would derive more information from utterance-final syllables than Mandarin listeners.

5.2.1 Participants

Sixty native listeners (aged 18-35) participated in the perception study: twenty from each of English, Cantonese, and Mandarin. The native English listeners were born and raised in Canada, except for two listeners who moved to Canada at the age of four or later. Among the native Cantonese listeners, seven originated from Hong Kong, six from Guangdong (China), and six from Canada. The native Mandarin listeners were from different regions of China, excluding Hong Kong. Table 5.1 shows the demographic details for these listeners.

Table 5.1. Demographics of the listeners in the perception study.

Language	Number of		Age (years)		Age Range (years)		
	Male	Female	Mean	SD	18-23	24-29	30-35
English	10	10	23.30	5.73	13	4	3
Cantonese	10	10	23.10	3.74	13	5	2
Mandarin	10	10	24.95	3.07	7	11	2

Ten other listeners also participated in this study: four English, three Cantonese, and three Mandarin. Their data were not used for the following reasons. Of the English listeners, two did not complete the study and two did not follow the instructions properly. Of the Cantonese listeners, one was non-native and two were over age 35. Of the Mandarin listeners, one was non-native and two moved from China to Canada before age ten.

These participants were recruited from the Introduction to Linguistics course or from flyers posted at the University of Calgary. They were fluent in speaking and listening to their native languages, and reported no visual, speech, or hearing impairments. The Cantonese and Mandarin participants were also fluent in reading and understanding English. For their participation in both sessions of the perception study, each listener received 2% course credit or $30.

5.2.2 Stimuli

For each language, the stimuli comprised 20 target pairs of sentences from four randomly selected speakers, totaling 160 sentences (5 blocks x 4 dialogues x 2 sentence types x 2 speakers x 2 genders). These sentences were gated into five different stimulus types, as shown in Table 5.2, resulting in a total of 800 tokens.

Table 5.2. Stimulus types.

Stimulus Type	Description	Example
Whole	Whole utterance	Mary is buying a chair
NoLast	All but the last syllable	Mary is buying a
Last	Last syllable	chair
Last2	Last two syllables	a chair
First	Utterance-initial name	Mary

Originally, I included stimulus type NoLast2 (‘all but the last two syllables’) in the experimental design to provide symmetry with stimulus type Last2. After pilot testing the design, I decided to replace it with stimulus type First. I made this decision because all of the utterance-initial names consisted of two syllables (except for ‘Ann’ in the 5-syllable

English sentences[13]), and in Mandarin (as mentioned in Section 4.2.2), these two syllables in blocks A to D carried pairs of tones that revealed the relative extent of the speaker's pitch range in the utterances. I wanted to find out if the pitch range in just these names could serve as a cue to help Mandarin listeners to differentiate between statements and questions.

This modified gating method differs from the typical method of gating stimuli, in which fragments of the utterance are increased in duration incrementally, for example, every 40 milliseconds (Grosjean, 1980). A 10% increment would parallel conditions 2-11 of the MpC method, described in Section 4.2.5. However, this approach would create too many test tokens for the identification task (40 sentences x 10 gated fragments x 4 speakers = 1600 tokens). Since the logistic regression in Section 4.3.2 showed that correct classification based on F0 cues, on average, is highest when the MpC model is presented with the initial 70-100% duration of the utterances and lowest when presented with the initial 0-10% duration, I decided that it would be more effective and efficient to focus on these parts of the utterances. Also, I was interested in the effects of the sentence-final stress/tone on the perception of sentence intonation. Since both stress and tones are associated with syllables, the utterances were gated at syllable boundaries.

To create the five stimulus types, I marked their boundaries in the textgrids of the original sound files in Praat (Boersma & Weenink, 2013), as shown in Figure 5.1. Then I wrote a Praat script, which extracted these five types of stimuli from each target sound file. The intensity across all stimuli was normalized to 65 decibels (dB) to avoid bias due

[13] The utterance-initial name was monosyllabic for the English sentences that were 5 syllables long because these sentences had too few syllables to fit a disyllabic name while keeping the sentence meaning similar to the meaning of the 5-syllable sentences in Cantonese and Mandarin.

to differences in loudness. The sound pressure level of conversation is approximately 60 dB (Raphael, Borden, & Harris, 2011: 33).

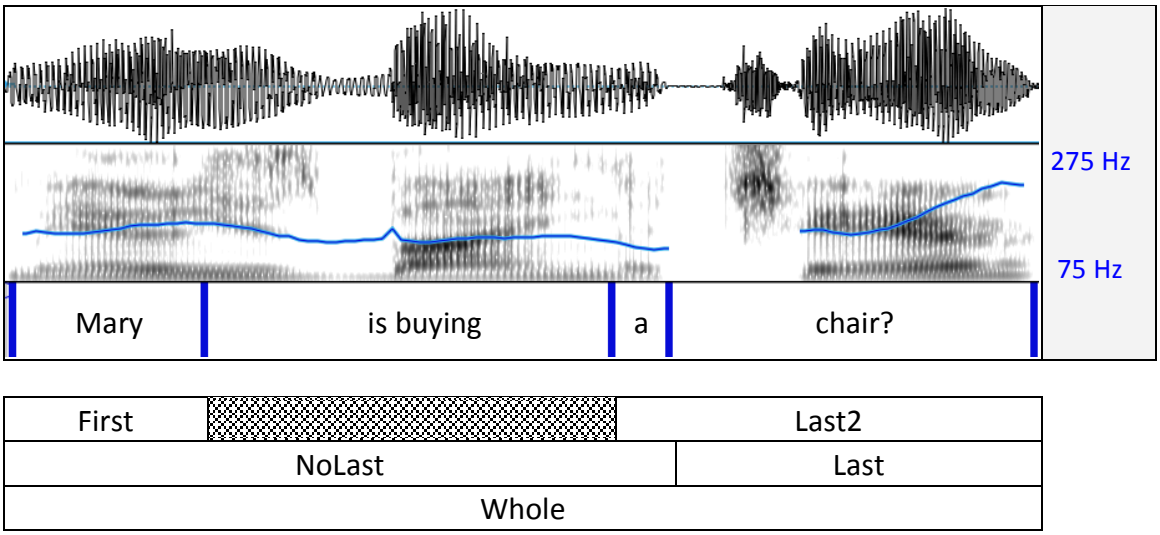


Figure 5.1. A marked textgrid for segmenting into the five stimulus types.

In order to determine the effects of the F0 differences between statements and questions on listeners' identification of these sentence types, I analyzed the F0 contours of statements and questions. For each stimulus, I extracted[14] the F0 values from eleven equidistant time points, as described in Section 3.3. Then I calculated characteristic statement and question contours using the mean F0 at each point for all four speakers. Figures 5.2-5.3 show the mean F0 contours, grouped by stimulus type and language.

[14] When a portion of an utterance was creaky voiced or voiceless, Praat failed to detect its F0 correctly, due to irregular voicing or lack of periodicity. For these portions of the utterances, I manually corrected their F0s in Praat using the Manipulation function before extracting their F0 values.

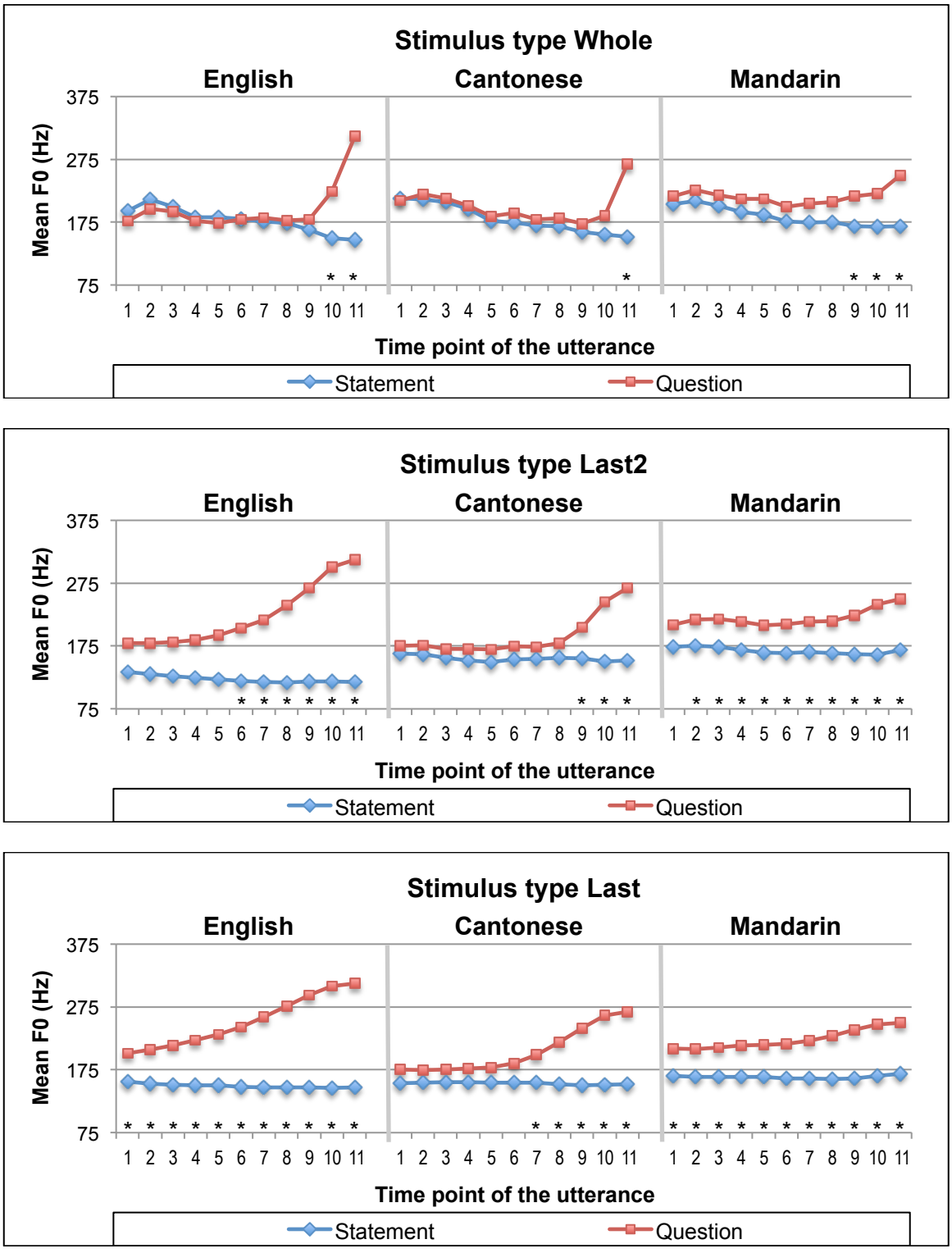


Figure 5.2. F0 contours of stimulus types Whole, Last2, and Last (* $p < 0.001$).

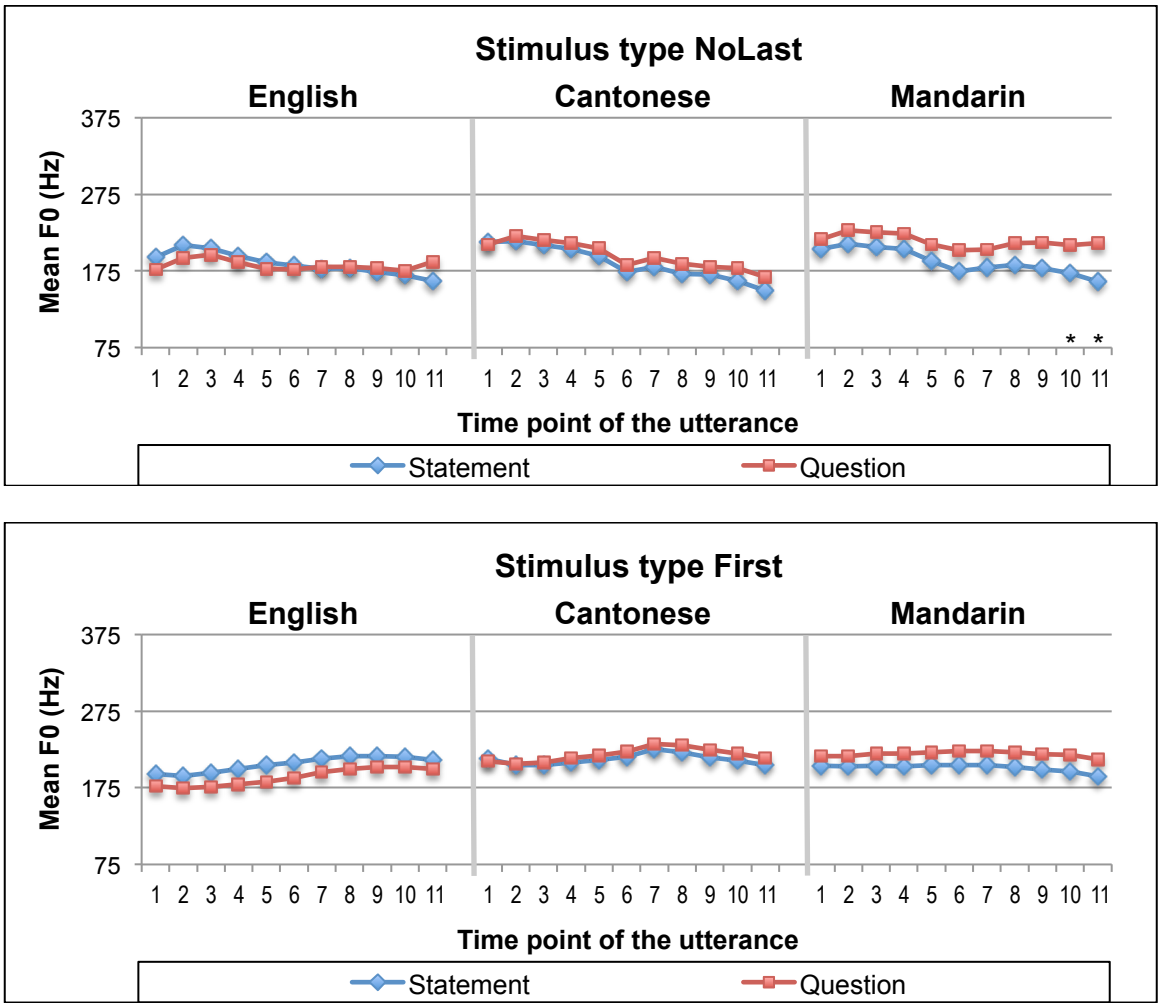


Figure 5.3. F0 contours of stimulus types NoLast and First (* $p < 0.001$).

For each group, I ran a repeated measures ANOVA with F0 as the dependent measure and with sentence type and time point as independent factors. At $\alpha = 0.001$, significant interactions between time point and F0 were found for all stimulus types [$F > 4.9$, $p < 0.001$], except for stimulus type First.[15] Post-hoc Tukey HSD tests revealed significant F0 differences between statements and questions at the time points indicated with an asterisk in Figures 5.2-5.3. Overall, these F0 differences extended through more of the stimulus types (Whole, Last2, and Last) that included the final syllable than in the stimulus types

[15] I used a smaller α-criterion here ($\alpha = 0.001$) than elsewhere ($\alpha = 0.05$) to focus on the bigger and more consistent differences.

(NoLast and First) that excluded the final syllable. In addition, significant F0 differences extended through more of stimulus type Last (English = 100%; Cantonese ≈ 45%; Mandarin = 100%), than in Last2 (English ≈ 55%; Cantonese ≈ 25%; Mandarin ≈ 95%), and through more of Last2 than Whole (English ≈ 15%; Cantonese ≈ 5%; Mandarin ≈ 25%). Across all three languages, they were relatively longer in Mandarin than in English, and in English than in Cantonese. These F0 differences reflect their language-specific intonation patterns: the question rise starts at the beginning of the utterance in Mandarin, near the nuclear tone in English, and in the final syllable in Cantonese.

In order to investigate the differences in F0 between final stressed and unstressed syllables and across all final tones, Figure 5.4 groups the F0 contours of stimulus type Last (and also Last2 for English) by stress and tone. For each of these subgroups, I once again ran a repeated measures ANOVA with F0 as the dependent measure and with sentence type and time point as independent factors. At $\alpha = 0.001$, all of these ANOVAs revealed significant interactions between sentence type and time point [$F > 5.2$, $p < 0.001$]. Post-hoc Tukey HSD tests revealed significant F0 differences between statements and questions at the time points indicated with an asterisk in Figure 5.4. In English, these significant differences in F0 extended through more of the final unstressed syllable (100%) than in the final stressed syllable (≈ 55%). In Cantonese, the duration of these significant F0 differences was relatively similar across the final tones (≈ 25-35%). In Mandarin, however, the duration of these F0 differences varied across all four final tones (high = 0%; rising ≈ 25%; low ≈ 15%; falling = 100%).

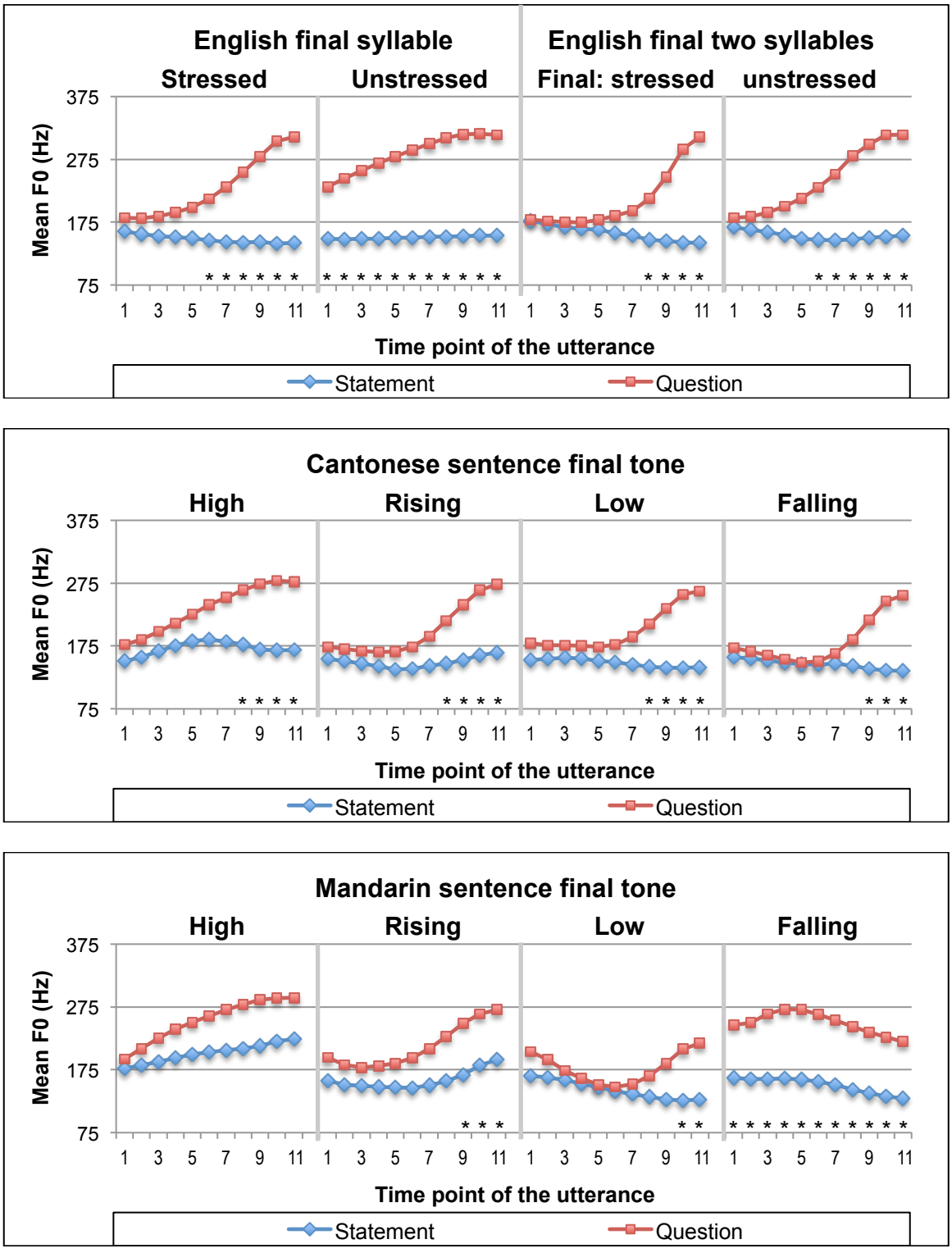


Figure 5.4. Final stress or tone in English, Cantonese, and Mandarin (* $p < 0.001$).

Therefore, in terms of identifying statements and questions accurately based on their F0 differences, the patterns in Figure 5.2 suggest Mandarin > English > Cantonese, whereas the patterns in Figure 5.4 suggest English > Cantonese > Mandarin.

5.2.3 Identification Task

The identification task was created in SuperLab 5, a computer application. In session 1, it consisted of five parts (Parts I-V): two practice phases, one training phase, and two testing phases. In session 2, it included two other parts (Parts VI-VII): an additional practice phase and an additional testing phase. The identification task in session 1 was shorter because the listeners had an extra task of filling out a questionnaire. Table 5.3 lists the number of trials and the types of stimuli presented in each part.

Table 5.3. The stimulus type(s) and number of trials presented in each part.

Session	Part	Phase	Number of Trials	Stimulus Types
1, 2	I	Practice	4	Whole
1, 2	II	Training	80	Whole
1, 2	III	Testing	80	Whole
1, 2	IV	Practice	8	NoLast, Last
1, 2	V	Testing	160	NoLast, Last
2	VI	Practice	8	First, Last2
2	VII	Testing	160	First, Last2

To enable direct comparison between the native listeners and the exemplar-based model on their performances in the identification tasks, the stimuli and test conditions for both the listeners and the model were designed to be as similar as possible. Since I decided to implement a two-fold cross-validation for training and testing the model (Section 3.4), a two-fold cross-validation had to be used for training and testing the native listeners. First of all, the target pairs of sentences were arranged randomly in five

different orders. For each order, the first 50% of the stimuli (the first 40 pairs of the 80 ordered pairs of sentences, e.g., 1a) was used for training and the second 50% (the last 40 pairs, e.g., 1b) was used for testing in session 1. Then the stimuli for training and testing were reversed between session 1 (run #1 of the two-fold cross-validation) and session 2 (run #2). In addition, for each of the five random orders, the training and test stimuli were counterbalanced across listeners, generating ten listener orders. Table 5.4 lists the random orders, stimulus sets for training and testing, and listener orders, using their (alpha)numeric identifiers.

Table 5.4. Ten listener orders, generated from five random orders of the stimuli.

Random Order	Stimuli		Listener Order	Session 1		Session 2	
	1st 50%	2nd 50%		Training	Testing	Training	Testing
1	1a	1b	1	1a	1b	1b	1a
			2	1b	1a	1a	1b
3	3a	3b	3	3a	3b	3b	3a
			4	3b	3a	3a	3b
5	5a	5b	5	5a	5b	5b	5a
			6	5b	5a	5a	5b
7	7a	7b	7	7a	7b	7b	7a
			8	7b	7a	7a	7b
9	9a	9b	9	9a	9b	9b	9a
			10	9b	9a	9a	9b

5.2.4 Procedure

The listeners attended two one-hour sessions, one to seven days apart. On day one, they first completed a brief questionnaire on their language background (see Appendix B). On both days, they performed the identification task described in Section 5.2.3. For each language, one listener from each gender group was assigned to each of the ten different listener orders listed in Table 5.4.

The listeners were tested individually in the Perception Room of the Phonetics Lab at the University of Calgary. This private room, equipped with two iMac computers of the same model and year, allowed for testing of up to two participants at the same time. Using the same type of computers was important for measuring reaction time. It helped to ensure that the lag time between the key pressed and the information sent to the SuperLab application was consistent between the computers.

During the task, the listeners sat in front of one of the iMac computers and listened to the audio stimuli through circumaural headphones. In each phase (Table 5.3), the computer presented the randomized stimuli to them one at a time. The listeners were instructed to rate whether the stimulus they had just heard was a statement or a question by pressing the appropriate keys on the keyboard: 1 = definitely a statement, 2 = likely a statement, 3 = maybe a statement, 7 = maybe a question, 8 = likely a question, and 9 = definitely a question (as shown in Figure 5.5).

1	2	3		7	8	9
definitely	likely	maybe		maybe	likely	definitely
statement				question		

Figure 5.5. Numerical keys corresponding to the gradient and categorical responses.

The purpose of using gradient ratings for each category was to determine if there was a correlation between the number of correct responses and the listeners' level of certainty in their responses (e.g., to determine how many incorrect responses were guesses). To reduce delays in pressing a response key due to eye or hand movements, the listeners were asked to keep their fingers on the response keys throughout the exercise. The reaction time was measured from the offset of the presentation of the stimulus to the onset of the pressing of a valid response key. Thus, the listeners were told to only take a

break (if they needed one) during a prompted break or at the end of a phase. Finally, to elicit low-level perceptual reactions, listeners were asked to respond as fast and accurately as they could.

The training phase provided immediate feedback to the listener, after each trial, on the type of sentence (statement or question) that had been presented in the trial. The testing phase provided feedback after every ten trials on just the number of correct responses the listener had provided in those ten trials in order to help maintain the listener's attention. At the end of each phase, the computer program also reported the total score of correct responses that the listener had registered during the entire phase. (Some listeners commented afterwards that the feedback on the scores made the task more interesting.) For consistency in the number of times that listeners heard the stimuli, the trials were presented only once, but there was no time limit on responding. Typically, the listeners completed each session in 25-45 minutes.

The practice exercise was designed to familiarize the listeners on how to do the task; the results from this phase were not analyzed. In each practice phase, listeners heard the same target pairs of sentences from two practice dialogues produced by a speaker whose utterances were not used in training and testing. The training phase, on the other hand, was intended to create a baseline set of experiences among all listeners of the same language. The assumption (based on exemplar theory principles) was that, prior to testing, each listener would have categorized in memory the same number of sentences produced by the speakers in the training phase. Despite it being merely a training exercise, the listeners were asked to rate the utterances first before being told what the sentence type was because 1) it turned training from a passive into an active exercise to

increase their attention, 2) this provided data on whether listener performance improved with each trial, and 3) it provided information on the individual listeners' baseline scores on the same task that would be used in testing.[16]

5.2.5 Statistical Analysis

Data from the perception study were analyzed in R (Urbanek & Iacus, 2013) for 1) perceptual sensitivity, 2) response bias, and 3) normalized reaction time.[17] Signal detection theory (Macmillan & Creelman, 2005) was used to measure perceptual sensitivity (d-prime) and response bias (beta) from z-scores (z) based on the percentage of correct 'statement' responses (hits) and incorrect 'statement' responses (false alarms), as shown in Table 5.5.

Table 5.5. Application of signal detection theory to 'statement' and 'question' responses.

Utterance	*Listener's Response*	
	STATEMENT	QUESTION
STATEMENT	Hit	Miss
QUESTION	False Alarm (FA)	Correct Rejection

D-prime (d') is the normalized difference between the proportions of hits and false alarms in a confusion matrix, as in (5.1). The higher the d-prime, the more sensitive the listeners are to the distinction between statements and questions in the signal. Beta (ß) is defined as the negative sum of hits and false alarms divided by two, as in (5.2).

(5.1) $d' = z(\%\ \text{Hit}) - z(\%\ \text{FA})$

(5.2) $ß = -1/2 * (z(\%\ \text{Hit}) + z(\%\ \text{FA}))$

[16] The listeners' performance during training was not analyzed for this thesis.

[17] The effect of listeners' ratings on percent scores was also analyzed. The preliminary results from an ANOVA revealed significant interactions between stimulus type NoLast and ratings on percent scores for English and Mandarin, and between stimulus type Last and ratings on scores for Mandarin. Further analysis was necessary but is beyond the scope of this thesis.

A positive beta means a bias in the listeners towards providing more 'statement' responses (independently of the signal), whereas a negative beta means a bias towards 'question' responses.

I also analyzed listener reaction time (RT) to find out if listeners identified a particular sentence type or stimulus type faster than another, and to determine if there were cross-linguistic differences in the timecourse of the responses. To normalize RTs across listeners, I converted RTs from milliseconds to z-scores, using the formula in (5.3) as follows: 1) I considered only RTs for correct responses to stimuli, 2) I calculated for each listener his or her average RT and the standard deviation of RT for these correct responses, and 3) I calculated a z-score for each correct RT, based on these listener-specific mean and standard deviation values.

(5.3) normalized_RT = (correct_response's_RT – mean_RT) / SD_RT

5.3 Results

In a preliminary analysis, pairwise t-tests showed no significant difference between male and female listeners on d' and ß (at $\alpha = 0.05$). Therefore, data from both genders were analyzed together in the analyses that follow. As mentioned in Section 5.2.3, the training and test stimuli were reversed between run #1 and run #2 in the two-fold cross-validation. In other words, the listeners had already heard the test stimuli (of type Whole only) in session 2 from the session 1 training phase. To determine the effect of this previous experience on listeners' performance, I ran additional pairwise t-tests. No significant difference was found between session 1 and session 2 on d' and ß for the stimulus types that had been presented in both of these sessions: Whole, NoLast, and Last. Therefore,

only data from session 2 will be analyzed from here on, since that set of data contained all five stimulus types. Finally, in aligning with the goals outlined in Section 5.1, the analysis of the results of the statistical tests will mainly focus on the significant interactions between stimulus type and language and between stimulus type and stress/tone by language.

5.3.1 Perceptual Sensitivity: Across Languages

To compare how well listeners across all three languages performed in the identification task, I ran a two-way analysis of variance (ANOVA) with d' as the dependent measure and with language and stimulus type as independent factors. The results showed a significant main effect of language [$F(2, 285) = 49.0$, $p < 0.001$] and stimulus type [$F(4, 285) = 767.8$, $p < 0.001$]. There was also a significant interaction between language and stimulus type [$F(8, 285) = 26.7$, $p < 0.001$].

A post-hoc Tukey HSD test revealed the following results. The English listeners ($\bar{x} = 2.85$) performed significantly better than the Cantonese listeners ($\bar{x} = 2.50$), and they both performed significantly better than the Mandarin listeners ($\bar{x} = 2.25$). Overall, the listeners' performance on the stimulus types, from best to worst, were as follows: Whole ($\bar{x} = 4.06$) > Last2 ($\bar{x} = 3.64$) > Last ($\bar{x} = 2.99$) > NoLast ($\bar{x} = 1.72$) > First ($\bar{x} = 0.27$). Figure 5.6 displays the interaction between language and stimulus type on d'.[18]

[18] Significant differences found between levels of a factor are indicated by '*' for the level(s) with the higher mean value(s) and '+' for the level(s) with the lower mean value(s) in the graphs reported in this thesis.

Both the English and Cantonese listeners performed significantly better than the Mandarin listeners on stimulus types that included the final syllable (Whole, Last2, and Last). The English listeners also performed significantly better than the Cantonese listeners on stimulus type Last. Both the English listeners and Mandarin listeners performed significantly better than the Cantonese listeners on stimulus type NoLast. In addition, the Mandarin listeners performed significantly better than the English listeners on stimulus type First.

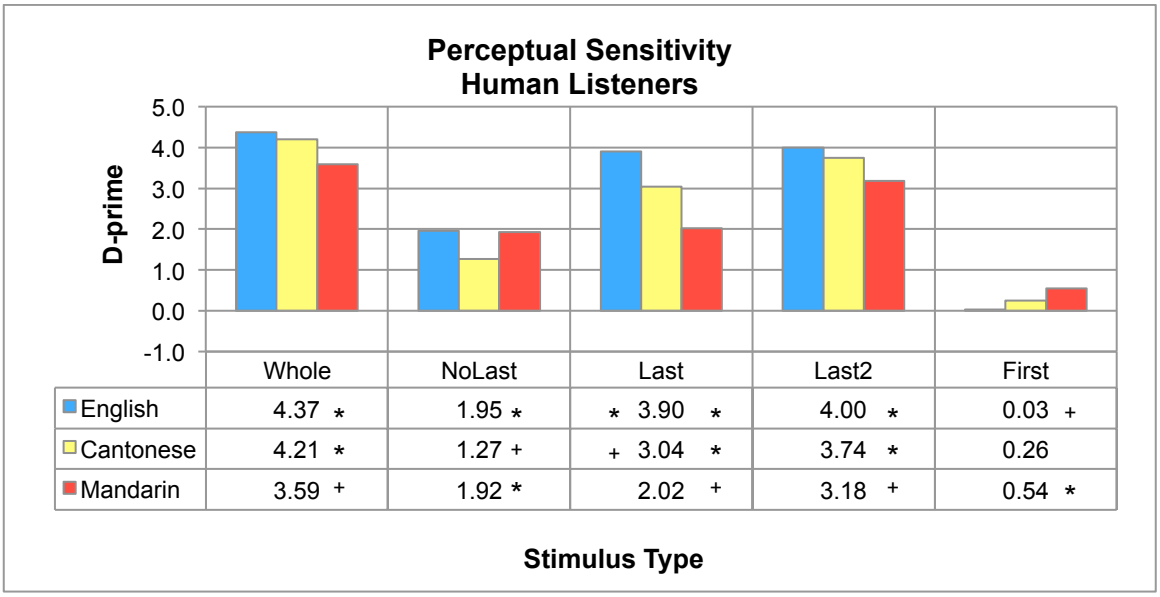


Figure 5.6. Interaction between language and stimulus type on d'.

In general, the listeners in each language performed significantly better on stimulus types that included the final syllable (Whole, Last2, and Last) than on stimulus types that excluded it (NoLast and First), as shown in Table 5.6.

Table 5.6. Interaction between language and stimulus type on d'.

Language	**Stimulus Types** (ordered by mean d' scores)
English	Whole, Last2, Last > NoLast > First; Whole > Last
Cantonese	Whole, Last2 > Last > NoLast > First
Mandarin	Whole, Last2 > Last, NoLast > First

5.3.2 Perceptual Sensitivity: Effect of Final Stress/Tone

To determine how the suprasegmental structures of stress and tone affected the listeners' performance, I ran a two-way ANOVA for each language, with d' as the dependent measure and with stimulus type and stress/tone as independent factors.[19] Significant interactions between stimulus type and stress/tone were found in English [$F(4, 190) = 25.1$, $p < 0.001$], Cantonese [$F(12, 380) = 2.9$, $p < 0.001$], and Mandarin [$F(12, 380) = 2.9$, $p < 0.001$].

Post-hoc Tukey HSD tests revealed the following results, as indicated in Figures 5.7-5.9. In English, for stimulus type NoLast, the listeners performed significantly better when the missing syllable was unstressed than when the syllable was stressed (Figure 5.7).

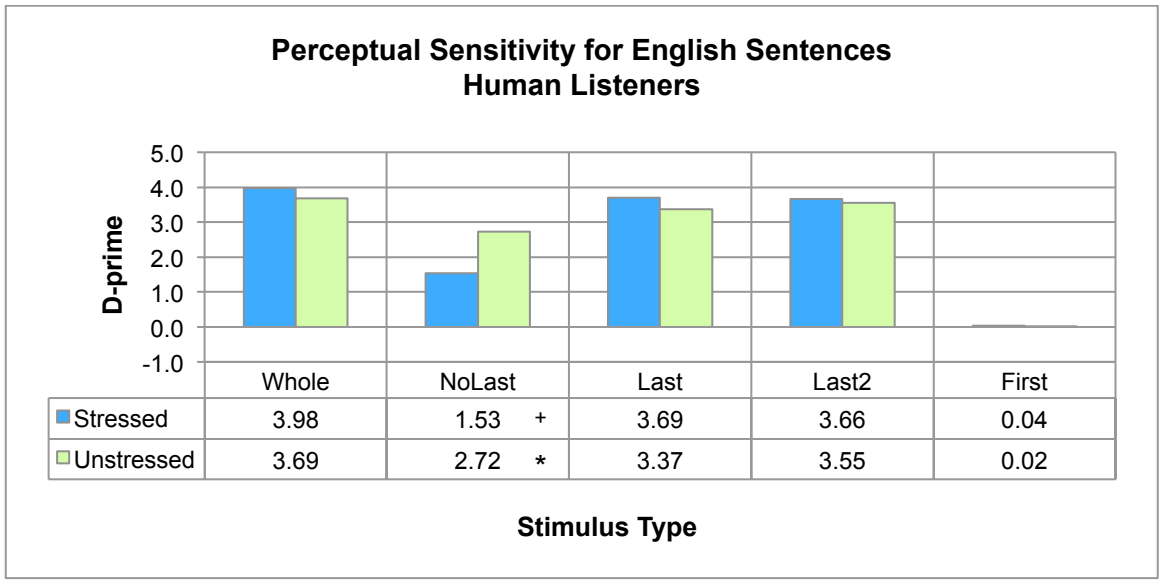
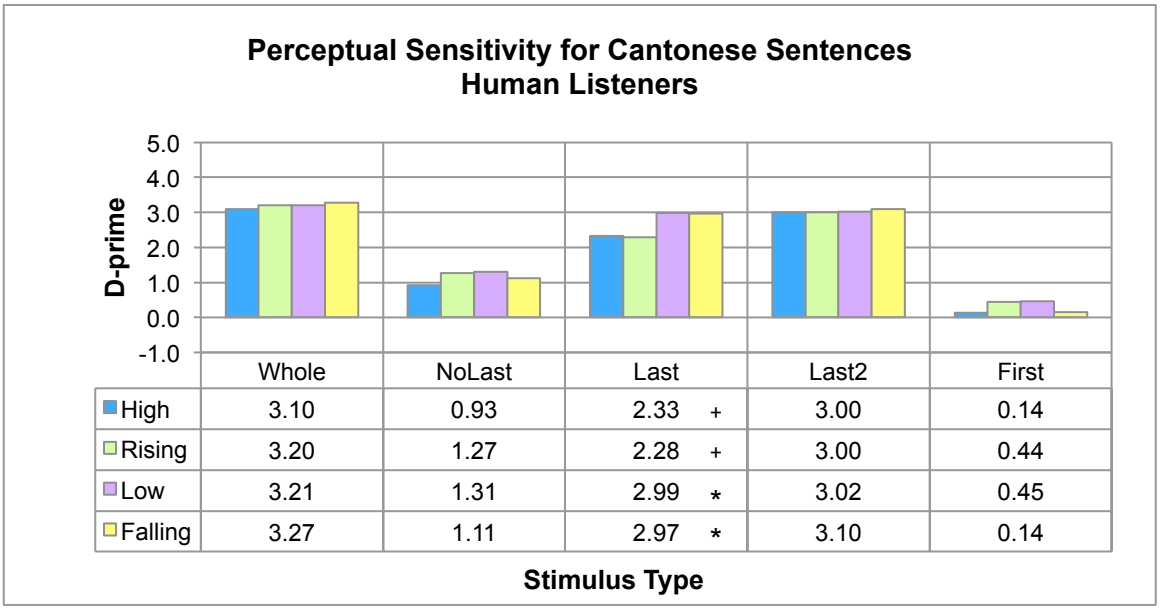


Figure 5.7. Interaction between stimulus type and stress on d' for English.

[19] Since the stress and tone categories were different for each language, I was unable to run one three-way ANOVA with language, stimulus type, and stress/tone as independent factors instead.

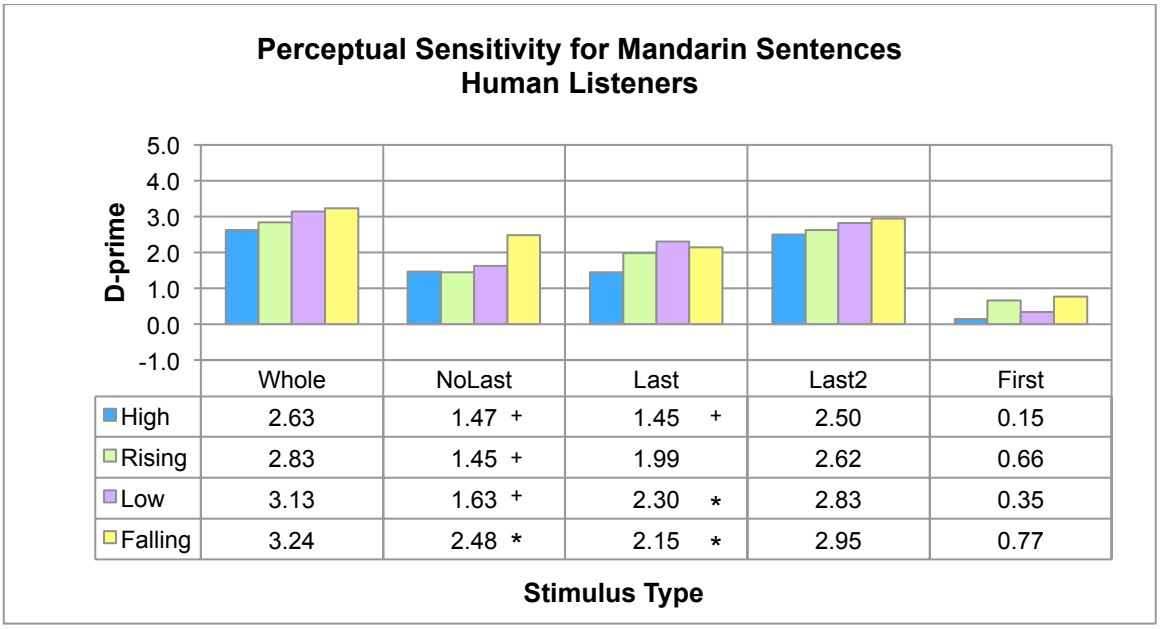
In Cantonese, on stimulus type Last, the listeners performed significantly better when the syllable ended in a low or falling tone than in a high or rising tone (Figure 5.8).



	Whole	NoLast	Last		Last2	First
High	3.10	0.93	2.33	+	3.00	0.14
Rising	3.20	1.27	2.28	+	3.00	0.44
Low	3.21	1.31	2.99	*	3.02	0.45
Falling	3.27	1.11	2.97	*	3.10	0.14

Figure 5.8. Interaction between stimulus type and tone on d' for Cantonese.

In Mandarin, on stimulus type NoLast, the listeners performed significantly better when the missing syllable carried a falling tone than a high, rising, or low tone (Figure 5.9). On stimulus type Last, these listeners performed significantly better when the syllable carried a low or falling tone than a high tone.



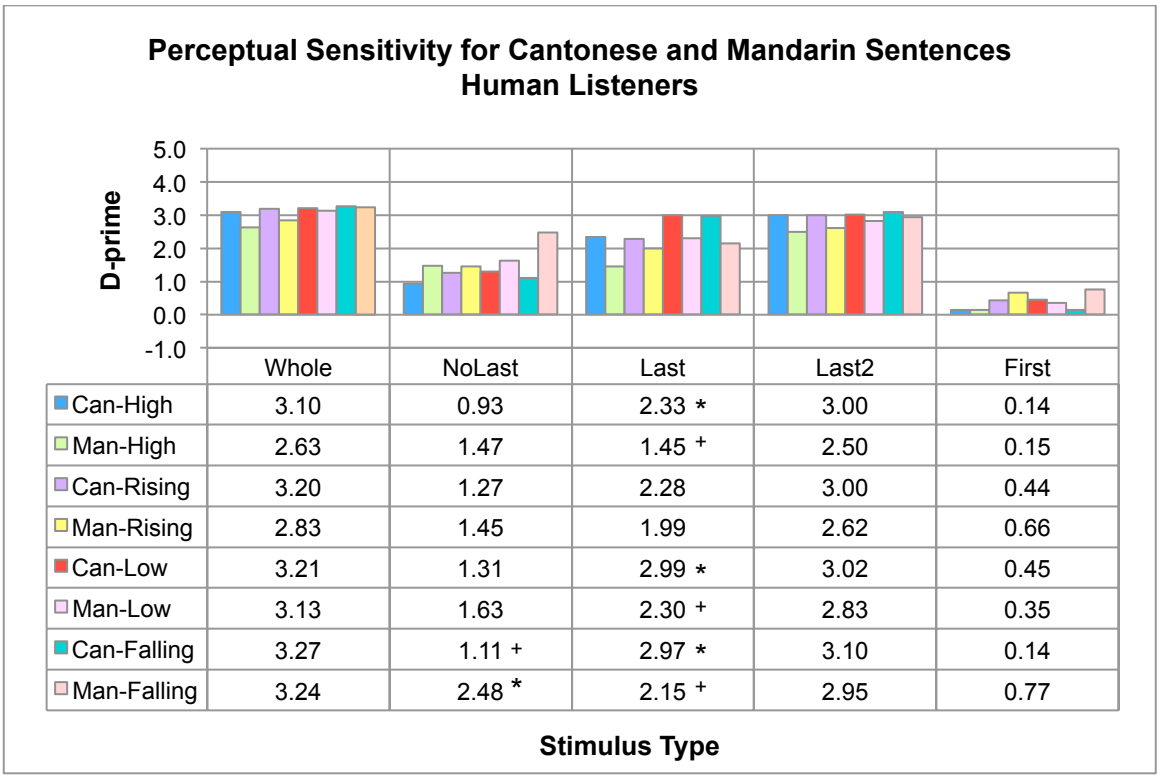
	Whole	NoLast	Last	Last2	First
High	2.63	1.47 +	1.45 +	2.50	0.15
Rising	2.83	1.45 +	1.99	2.62	0.66
Low	3.13	1.63 +	2.30 *	2.83	0.35
Falling	3.24	2.48 *	2.15 *	2.95	0.77

Figure 5.9. Interaction between stimulus type and tone on d' for Mandarin.

5.3.3 Perceptual Sensitivity: Between Tone Languages

To compare how the suprasegmental structure of tone affected the Cantonese and Mandarin listeners' performance in the identification task, I ran a three-way ANOVA to analyze the combined Cantonese and Mandarin data set, with d' as the dependent measure and with language, stimulus type, and tone as independent factors. A significant interaction among language, stimulus type, and tone was found [$F(12, 760) = 3.0$, $p < 0.001$].

Figure 5.10 displays the significant differences among these three factors that were revealed by a post-hoc Tukey HSD test. On stimulus type NoLast, the Mandarin listeners performed significantly better than the Cantonese listeners when the missing syllable carried a falling tone. On stimulus type Last, the Cantonese listeners performed significantly better than the Mandarin listeners when the syllable carried a high, low, or falling tone.



	Whole	NoLast	Last	Last2	First
Can-High	3.10	0.93	2.33 *	3.00	0.14
Man-High	2.63	1.47	1.45 +	2.50	0.15
Can-Rising	3.20	1.27	2.28	3.00	0.44
Man-Rising	2.83	1.45	1.99	2.62	0.66
Can-Low	3.21	1.31	2.99 *	3.02	0.45
Man-Low	3.13	1.63	2.30 +	2.83	0.35
Can-Falling	3.27	1.11 +	2.97 *	3.10	0.14
Man-Falling	3.24	2.48 *	2.15 +	2.95	0.77

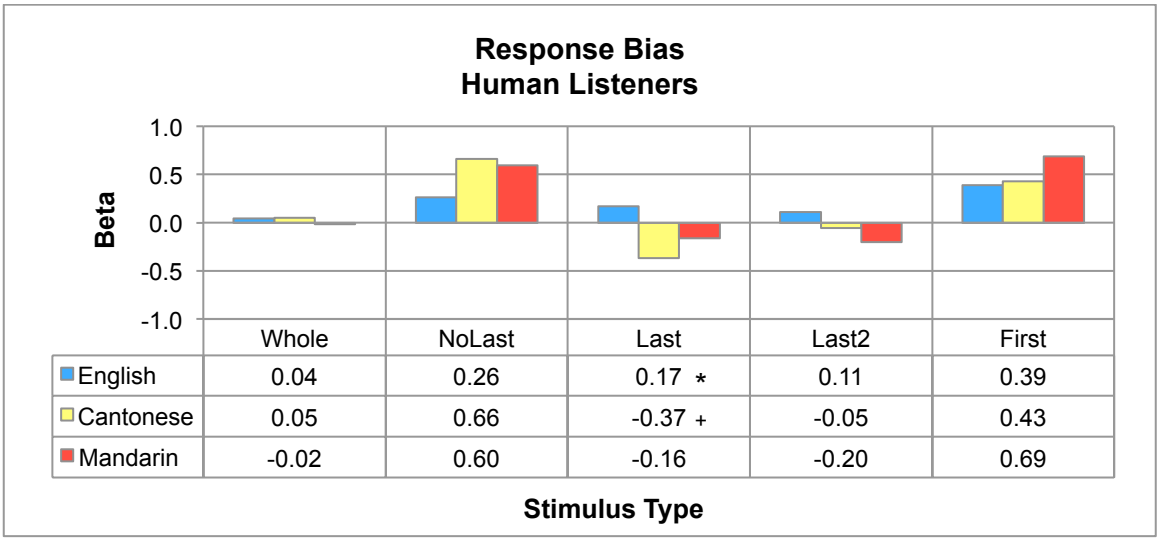
Figure 5.10. Interaction among language, stimulus type, and tone on d'.

5.3.4 Response Bias: Across Languages

To determine whether the listeners responded to the stimuli in the identification task with particular biases towards statements or questions, I ran a two-way ANOVA with ß as the dependent measure and with language and stimulus type as independent factors. At $\alpha =$

0.05, there was no significant main effect of language, but there was a significant main effect of stimulus type [$F(4, 285) = 34.0$, $p < 0.001$]. There was also a significant interaction between language and stimulus type [$F(8, 285) = 5.1$, $p < 0.001$].

A post-hoc Tukey HSD test revealed the following results. There was significantly more bias towards statements than questions in the listeners' responses to stimulus types NoLast ($\bar{x} = 0.50$) and First ($\bar{x} = 0.50$) than in the listeners' responses to stimulus types Whole ($\bar{x} = 0.02$), Last ($\bar{x} = -0.12$), and Last2 ($\bar{x} = -0.05$). Figure 5.11 displays the interaction between language and stimulus type on ß. There was significantly more bias towards statements than questions in the English listeners' responses to stimulus type Last than in the Cantonese listeners' responses.



	Whole	NoLast	Last	Last2	First
English	0.04	0.26	0.17 *	0.11	0.39
Cantonese	0.05	0.66	-0.37 +	-0.05	0.43
Mandarin	-0.02	0.60	-0.16	-0.20	0.69

Figure 5.11. Interaction between language and stimulus type on ß.

In general, the Cantonese and Mandarin listeners responded with significantly more bias towards statements than questions to stimuli that excluded the utterance-final syllable than to stimuli that included the utterance-final syllable, as shown in Table 5.7. No significant bias was found in the English listeners' responses across stimulus types.

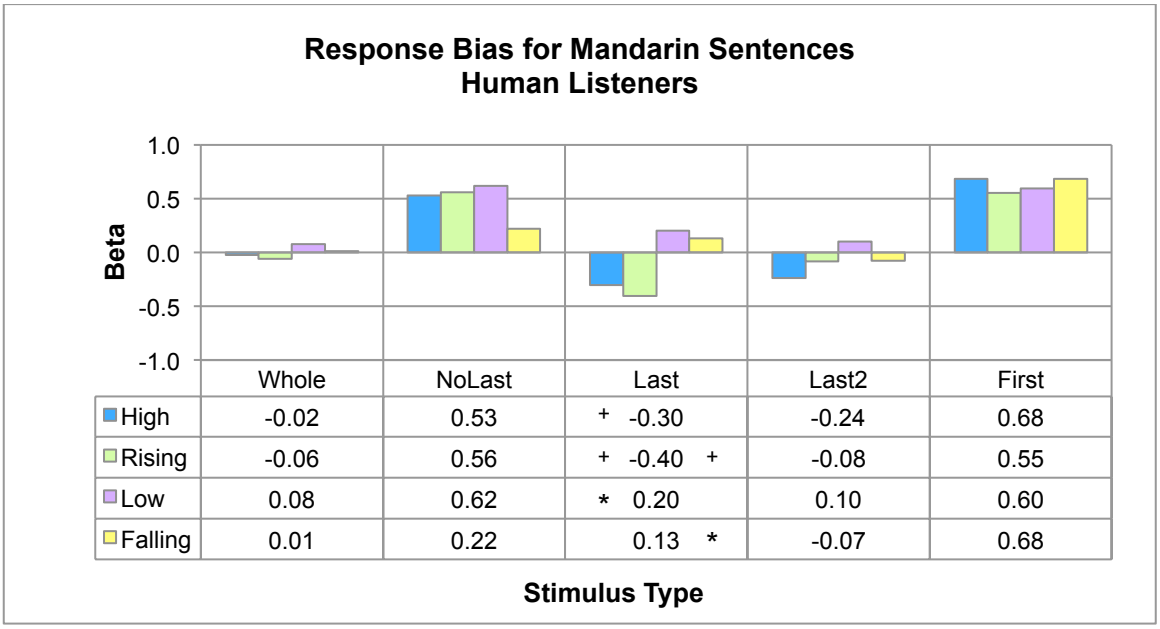
Table 5.7. Interaction between language and stimulus type on ß.

Language	**Stimulus Types** (ordered by mean ß values)
Cantonese	NoLast, First > Last2, Last; NoLast > Whole
Mandarin	NoLast, First > Whole, Last2, Last

5.3.5 Response Bias: Effect of Final Stress/Tone

To determine how the suprasegmental structures of stress and tone affected the listeners' behaviour in responding to the statement and question utterances in the identification task, I ran two-way ANOVAs for each language, with ß as the dependent measure and with stimulus type and stress/tone as independent factors. At $\alpha = 0.05$, there was a significant interaction between stimulus type and stress/tone for Mandarin [$F(12, 380) = 2.9$, $p < 0.001$], but not for English and Cantonese.

Post-hoc Tukey HSD tests revealed that, for stimulus type Last, the Mandarin listeners responded with significantly more bias towards statements to stimuli that carried a low tone than to stimuli that carried a high or rising tone, and to stimuli that carried a falling tone than to stimuli that carried a rising tone (Figure 5.12).



	Whole	NoLast	Last	Last2	First
High	-0.02	0.53	+ -0.30	-0.24	0.68
Rising	-0.06	0.56	+ -0.40 +	-0.08	0.55
Low	0.08	0.62	* 0.20	0.10	0.60
Falling	0.01	0.22	0.13 *	-0.07	0.68

Figure 5.12. Interaction between stimulus type and tone on ß for Mandarin.

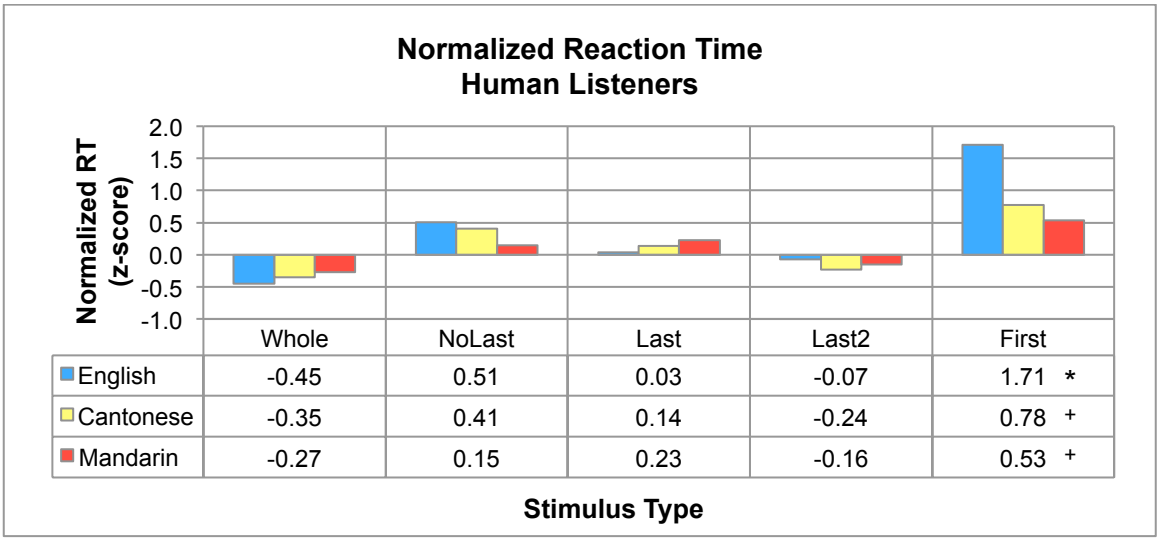
5.3.6 Response Bias: Between Tone Languages

To compare how the suprasegmental structure of tone affected the behaviour of the Cantonese and Mandarin listeners in responding to the statement and question utterances, I ran a three-way ANOVA on the combined Cantonese and Mandarin data set, with ß as the dependent measure and with language, stimulus type, and tone as independent factors. At $\alpha = 0.05$, there was neither a significant main effect of language (as Section 5.3.4 also found) nor a significant interaction among language, stimulus type, and tone.

5.3.7 Reaction Time: Across Languages

To investigate whether reaction time correlates with the listeners' familiarity with the intonation patterns of the languages, I ran a two-way ANOVA with normalized RT as a dependent measure and with language and stimulus type as independent factors. The results showed significant main effects of language [$F(2, 585) = 14.5$, $p < 0.001$] and stimulus type [$F(4, 585) = 137.5$, $p < 0.001$]. There was also a significant interaction between language and stimulus type [$F(8, 585) = 14.8$, $p < 0.001$].

A post-hoc Tukey HSD test revealed the following results. The English listeners ($\bar{x} = 0.35$) responded significantly slower (with longer normalized RT) than the Cantonese speakers ($\bar{x} = 0.15$) and the Mandarin listeners ($\bar{x} = 0.10$). Overall, the listeners' performance on the stimulus types, from slowest to fastest, were as follows: First ($\bar{x} = 1.01$) > NoLast ($\bar{x} = 0.35$) > Last ($\bar{x} = 0.13$) > Last2 ($\bar{x} = -0.15$), Whole ($\bar{x} = -0.36$). Figure 5.13 displays the interaction between language and stimulus type on normalized RT. The English listeners were significantly slower than the Cantonese and Mandarin listeners in responding to stimulus type First.



	Whole	NoLast	Last	Last2	First
English	-0.45	0.51	0.03	-0.07	1.71 *
Cantonese	-0.35	0.41	0.14	-0.24	0.78 +
Mandarin	-0.27	0.15	0.23	-0.16	0.53 +

Figure 5.13. Interaction between language and stimulus type on normalized RT.

In general, the listeners responded the slowest to stimulus type First and the fastest to stimulus type Whole, as shown in Table 5.8.

Table 5.8. Interaction between language and stimulus type on normalized RT.

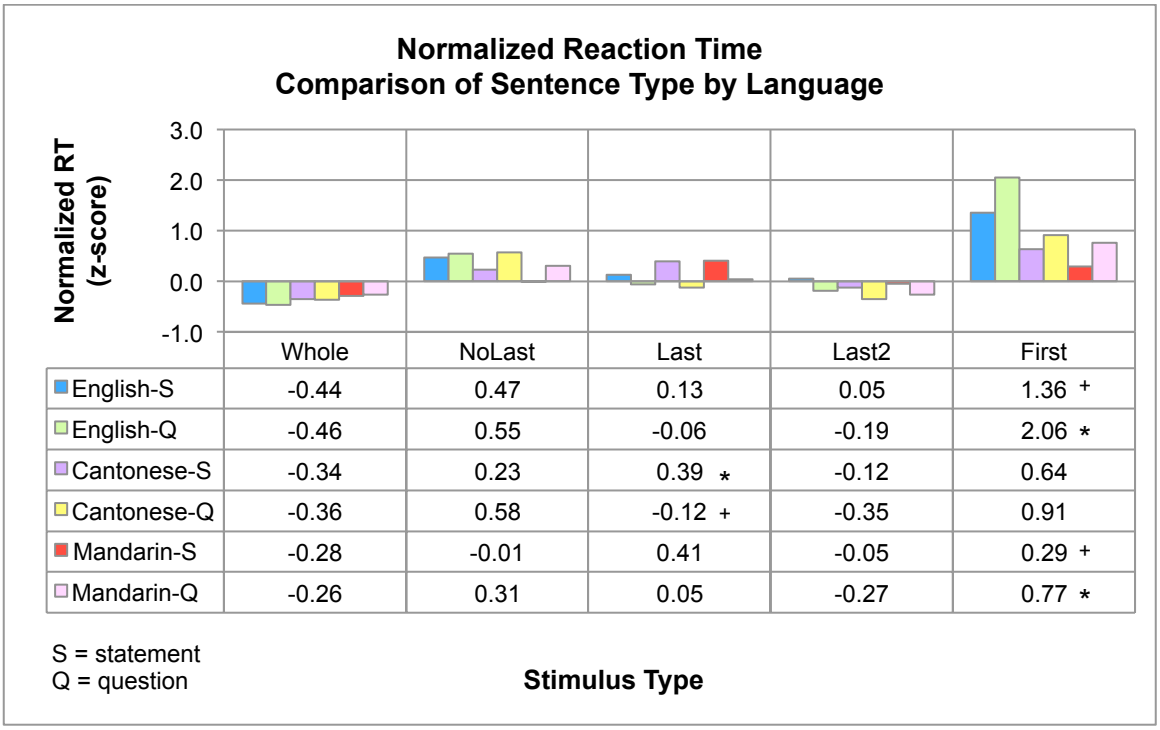
Language	**Stimulus Types** (ordered by mean normalized RTs)
English	First > NoLast > Last, Last2 > Whole
Cantonese	First > Last, Last2, Whole; NoLast > Last2, Whole; Last > Whole
Mandarin	First > NoLast, Last2, Whole; NoLast > Whole; Last > Last2, Whole

5.3.8 Reaction Time: Between Sentence Types

To compare how quickly the listeners responded to statement versus question utterances, I ran another two-way ANOVA for each language, with normalized RT as a dependent measure and with stimulus type and sentence type as independent factors.[20] This ANOVA revealed a significant interaction between stimulus type and sentence type for all three languages: English [$F(4, 190) = 4.3$, $p = 0.002$], Cantonese [$F(4, 190) = 6.4$, $p < 0.001$], and Mandarin [$F(4, 190) = 6.2$, $p < 0.001$].

[20] I also ran a three-way ANOVA with normalized RT as a dependent measure and with language, stimulus type, and sentence type as independent factors. There were significant main effects of language [$F(2, 570) = 15.9$, $p < 0.001$] and stimulus type [$F(4, 570) = 151.1$, $p < 0.001$] as well a significant interaction between language and stimulus type [$F(8, 570) = 16.2$, $p < 0.001$]. However, no significant interactions were found between language and sentence type and among language, stimulus type, and sentence type ($p > 0.05$).

Figure 5.14 displays the significant results revealed by a post-hoc Tukey HSD test. When presented with stimuli of type Last, the Cantonese listeners responded significantly slower to statement than to question utterances. When presented with stimuli of type First, the English and Mandarin listeners responded significantly slower to question than to statement utterances.



	Whole	NoLast	Last	Last2	First
English-S	-0.44	0.47	0.13	0.05	1.36 +
English-Q	-0.46	0.55	-0.06	-0.19	2.06 *
Cantonese-S	-0.34	0.23	0.39 *	-0.12	0.64
Cantonese-Q	-0.36	0.58	-0.12 +	-0.35	0.91
Mandarin-S	-0.28	-0.01	0.41	-0.05	0.29 +
Mandarin-Q	-0.26	0.31	0.05	-0.27	0.77 *

Figure 5.14. Interaction between stimulus type and sentence type on normalized RT.

5.4 Discussion

5.4.1 Cross-linguistic Performance

In the identification task, overall, the English listeners performed the best, followed by the Cantonese listeners, and then the Mandarin listeners. This ranking was anticipated by the duration of the significant F0 differences between statements and questions in the final syllable, as indicated in Figure 5.4. However, the duration of the significant F0

differences in Figures 5.2-5.3 suggested that Mandarin listeners would perform the best. Since Figures 5.2-5.3 showed the F0 contours with all final stress/tones combined, whereas Figure 5.4 showed the F0 contours by final stress/tone, this result provides evidence that the salient cue in distinguishing between statements and questions is in the final syllable and that the final stress/tone affects native listeners' perception of statement and question intonation. (Specific effects by final stress/tone are discussed in Section 5.4.3.) This analysis is supported by the fact that both the English and Cantonese listeners performed significantly better than the Mandarin listeners on stimuli that included the final syllable. As described in Section 2.6.2, English and Cantonese both signal questions with a pattern that includes a high boundary tone (H%), whereas Mandarin signals questions with a raised pitch range. The significantly higher d' scores for English and Cantonese, compared to Mandarin, suggest that the high F0 rise is a more salient acoustic cue than raised pitch range in identifying questions.

In addition, the English listeners performed significantly better than the Cantonese listeners on the final syllable alone. There are three possible reasons why this may have occurred. Firstly, although both English and Cantonese signal questions with a final F0 rise, the English question rise depends on the timing of the nuclear accent in the intonational phrase (Wells, 2006), whereas the Cantonese question rise is always at the right boundary of the intonational phrase (Wong et al., 2005). With the varied timing of the question rise in English, question rises which start earlier in the utterance have longer duration, providing more acoustic signal to the listener than rises which start later. Cantonese question rises start after the tone on the final syllable so, on average, they have a shorter duration than English question rises, as shown in Figure 5.2. The dependency on

H% in cueing Cantonese echo questions also explains why the English and Mandarin listeners performed significantly better than the Cantonese listeners on stimuli that excluded the final syllable. Secondly, the lexical tone on the final syllable can affect the perception of the question rise in Cantonese. Since Cantonese tones retain their F0 contours at the end of statements, a rising tone on the final syllable of a statement may be easily confused with a question rise. Thirdly, English statement intonation has a final fall, which increases the F0 gap between statements and questions towards the end of the utterance.

One final point—the Mandarin listeners performed significantly better than the English listeners on the first two syllables, likely due to the (slight) raise in pitch on Mandarin questions and the higher pitch in English statements than questions in this portion of the utterance, as shown in Figure 5.3.

In summary, a final high F0 rise in questions provides more discernible acoustic differences in discriminating between statements and questions than a global, raised pitch range in questions. Also, lexical tones have a more confusing effect on the perception of question intonation than does word stress. As well, listeners tend to associate higher pitch with questions and lower pitch with statements.

5.4.2 Performances Across Stimulus Types

In general, the listeners in the identification task performed significantly better when presented with stimuli which included the utterance-final syllable than stimuli which excluded it. This result supports the argument in Section 5.4.1 that the question rise is a salient cue in identifying statements and questions, since the final syllables in English and

Cantonese contain the question rise (and statement fall for English). In Mandarin, the final syllables of question utterances contain a final F0 rise as well (Liu & Xu, 2006) but one which is not as steep as the English rise and the Cantonese rise, as shown in Figure 5.2. In addition, among the two stimulus sets that included or excluded the final syllable, the listeners performed significantly better on the longer stimuli than on the shorter stimuli. The performance on whole utterances was significantly better than on the final syllables alone because 1) in English, the question rise can start before the final syllable, and 2) in Cantonese and Mandarin, the longer utterance provides contextual information, which helps to identify the tone of the final syllable. For similar reasons, performance on stimuli that lacked only the final syllable was significantly better than on stimuli that lacked more than just the final syllable because the penultimate syllable provides acoustic information for the question rise and/or context for the final tone, which helps the listener in performing the identification task.

Even though stimulus type Whole was the longest among the stimulus types, the listeners did not perform significantly better on it than on stimulus type Last2. This evidence suggests that the salient cues in identifying statements and questions are in not just the final syllable but in the combined final two syllables. It also explains why the Cantonese and Mandarin listeners performed significantly better on stimulus type Last2 than on Last. For English, there was no significant difference in performance found between stimulus types Last2 and Last. As mentioned earlier, the question rise in English can start in either the penultimate or the final syllable. However, in both cases, the maximum F0 of the rise occurs in the final syllable, as shown in Figure 5.2. The fact that the English listeners did not perform significantly better on stimuli consisting of the final

two syllables than on the final syllable alone suggests that these listeners focused on the difference in F0 height more than the duration of the F0 rise.

Finally, the fact that there was no significant difference in the Mandarin listeners' performance on stimulus type Last than NoLast suggests that although the slight question rise on the final syllable provides a perceptual cue to identify questions in Mandarin, it is no more a reliable cue than the raised pitch range in general.

In summary, the combined final two syllables provide the most salient acoustic cues for differentiating between statements and questions in English, Cantonese, and Mandarin. The acoustic information in the penultimate syllable, which may include the part of the question rise for English and the tonal contextual information for Cantonese and Mandarin, strengthens the question signal in the utterance.

5.4.3 Effects of Final Stress/Tone on Performance

In the identification task, stress or tone affected the listeners' responses to only the stimuli that included or excluded just the utterance-final syllable. In English, the listeners' performance on stimuli that excluded only the final syllable was significantly better when the missing syllable was unstressed than stressed because part of the question rise would still be available if the missing syllable was unstressed.

In Cantonese, the listeners' identification accuracy for the final-syllable stimuli was significantly better when it carried a low or falling tone than when it carried a high or rising tone, most likely because Cantonese statements retain the F0 direction of the final tone. In addition, high tones are often realized as rising tones due to transition from a preceding lower tone. Thus, statements which terminate with a high or rising tone are

easily confusable with questions. I have postulated above that the penultimate syllable provides tonal context which helps to identify the tone on the final syllable. This claim is further supported by the fact that final tone-based differences in performance disappeared for stimulus type Last2.

In Mandarin, the listeners' performance on stimuli that excluded only the final syllable was significantly better when the missing syllable was a falling tone than a high, rising, or low tone. A likely explanation for this effect is that the F0 difference between the statement and question at the end of stimulus type NoLast (which is the start of stimulus type Last) is much greater when the final syllable carried a falling tone than any of the other tones, as shown in Figure 5.4 at time point 1 of the Mandarin sentence-final tones. The Mandarin listeners also performed significantly better on the final syllable alone when the tone was low or falling than when the tone was high or rising. With the low tone, the stimuli may have been easier to identify because the final parts of the F0 contours of the statements and questions diverge in opposite directions (as shown at time point 7 of the Mandarin low-tone contour in Figure 5.4). With the falling tone, the tone is raised much higher relative to the rest of the tone in questions than in statements (as shown at time points 4 and 5 of the Mandarin falling-tone contour in Figure 5.4 and as indicated with %e-prom in Figure 2.9). That is, pitch direction may be a contributing factor. Additionally, in questions, the final low tone falls and then rises while the falling tone rises and then falls, but in statements, both tones fall. With the high and rising tones, the stimuli may have been harder to identify because both the statement and question contours rise towards the end (as shown in the Mandarin high-tone and rising-tone contours in Figure 5.4); there is no difference in pitch direction between both sentence

types. In an identification task between statements and questions involving 16 native listeners of Mandarin, Yuan and Shih (2004) also found that the native listeners identified questions ending in falling tones more accurately than questions ending in rising tones.

Between the two tone languages, Cantonese listeners performed significantly better than the Mandarin listeners on final syllables which ended in high, low, or falling tones, but not in rising tones. As mentioned earlier, the question cue in Cantonese is primarily on the final syllable and its statements with final rising tones are confusable with its questions. The Mandarin listeners, however, performed significantly better than the Cantonese listeners on stimuli that lacked the final syllable when the missing syllable ended in a falling tone. It is likely that the greater F0 difference between the statement and question at the right edge of the penultimate syllable of utterances missing the final falling tone in Mandarin helps the Mandarin listeners to identify these stimuli.

In summary, in English, stress affects the F0 information available from the penultimate syllable for identifying questions. In Cantonese and Mandarin, in general, final falling tones are less confusable than final rising tones in questions for the Cantonese and Mandarin listeners.

5.4.4 Listeners' Response Bias

Compared to stimuli that included the final syllable, the Cantonese and Mandarin listeners responded to stimuli that excluded the final syllable with significantly more 'statement' responses than 'question' responses. The fact that these listeners also reacted significantly slower to stimuli that excluded the final syllable than to stimuli that included this syllable suggests that these listeners tended to respond with 'statement' when they

were uncertain of the sentence type. On stimuli that comprised only the final syllable, the English listeners responded with more 'statement' responses than the Cantonese listeners. This response behaviour suggests that, for this particular stimulus type in which the question cue is present for these two languages, the English listeners' default response type is statement and they listen for the question cue, whereas the Cantonese listeners' default response type is question and they listen for the statement cue.

Comparing the different tones, the Mandarin listeners only responded with significantly more bias towards statements than questions to final syllables that ended in low or falling tones than they did to high and/or rising tones. This tendency echoes the commonly held notion that lower F0 and/or falling intonation are associated with statements, and higher F0 and/or rising intonation are associated with questions.

5.4.5 Reaction Time to the Intonation Cue

The normalized RTs of the listeners more or less reflected their performance patterns: they responded significantly faster to stimuli that included the utterance-final syllable than to stimuli that excluded it, suggesting that they recognized the salient, final-syllable question cue. Specifically, the Cantonese listeners responded significantly faster to question stimuli than to statement stimuli when presented with the final syllable alone, suggesting that they were cueing in on the high question rise. The English and Mandarin listeners, on the other hand, responded slower to question stimuli than to statement stimuli when presented with the first two syllables, suggesting that the question cues which they were listening for might have been obscured or missing in these stimuli. However, the Mandarin listeners did perform significantly better than the English

listeners on this stimulus type. Thus, the acoustic cues for identifying Mandarin questions were present in these stimuli, albeit somewhat weak. In fact, the question contour is slightly higher than the statement contour in Mandarin, as shown in Figure 5.3. In summary, listeners responded faster when the question intonation cue is in the stimuli and responded more slowly when it is missing.

Comparing all three languages reveals a positive relationship between the reaction times of the listeners (English > Cantonese > Mandarin) and their accuracy rates in identifying the sentence types (English > Cantonese > Mandarin): the longer the reaction time, the higher the accuracy rate. In general, this observed pattern is expected due to a trade-off between speed and accuracy. However, it is possible that the English listeners reacted more slowly because they required more time to process the following information: 1) 'heavy' syllables, 2) nonsense syllables, and 3) other acoustic information in addition to F0. First of all, the English syllables (e.g., C(C)VC(C)(C) such as [ˈhɪs.t͡ʃɹi] 'history' or [ˈfɪlmz] 'films') tend to be heavier than the Cantonese syllables (e.g., CV(C) such as [lik6 si:2] 'history') and the Mandarin syllables (e.g., CV such as [li4 ʂi3] 'history'). Secondly, half of the final words in the English utterances were polysyllabic (e.g., [ˌɛ̃n.dʒəˈnɪɹ] 'engineer'), which means that many of the fragmented utterances (i.e., stimuli of type NoLast, Last2, and Last) contained a nonsense syllable (e.g., [ˌɛ̃n.dʒə], [dʒəˈnɪɹ], or [ˈnɪɹ]). Contrastively, in Cantonese and Mandarin, a monosyllable usually has meaning (e.g., [lik6] or [li4] 'experience', and [si:2] or [ʂi3] 'chronicle'). Thirdly, English stress has more than one main acoustic correlate, including F0, duration, and intensity. The main acoustic correlate of Cantonese and Mandarin tones is F0.

Comparing all five stimulus types reveals a negative relationship between the reaction times of the listeners (First > NoLast > Last > Last2 > Whole) and their accuracy rates in identifying the sentence types (Whole > Last2 > Last > NoLast > First): the longer the reaction time, the lower the accuracy rate. In other words, the less sensitive the listener was to the statement/question distinction for one stimulus type than for another, the slower the listener responded to the former than to the latter. Also, this may reflect the amount of cognitive operations required to reconstruct the missing part of the stimulus.

In addition, the listener's response bias might have affected the reaction time as well. In particular, both stimulus types NoLast and First lacked the final syllable that contained the salient cue for questions. The listeners responded more slowly and with more bias towards statements than questions on these stimuli, as opposed to stimuli that included the final syllable. These listeners' bias suggests that they assumed an utterance was a statement until they found evidence for a question. Their slow response suggests that the processing time to determine the sentence type of an utterance fragment (e.g., to match up an utterance fragment with a whole utterance in memory) is longer for an utterance fragment than for a whole utterance, and for an utterance fragment that lacked the salient question cue than for an utterance fragment that contained this question cue.

5.5 Summary

This chapter has examined native listeners' perception of sentence intonation in English, Cantonese, and Mandarin, and established a baseline level of human performance in sentence-type identification in these three intonationally distinct languages. The next chapter describes a series of computer simulations by an exemplar-based model on the

same sentence-type identification task as these human listeners did, and reports on their comparative results.

Chapter 6: Perceptual Simulations of the Model

6.1 Goals

The goals of the computer simulations were 1) to test how accurately the proposed exemplar-based model could classify statements and questions in each language, based on F0 alone, 2) to compare how well the model could correctly classify statements and questions compared to the human listeners when both were tested with the same sets of stimuli, 3) to find out if the model's performance differed by language, and 4) to determine the effects of stress/tone on the classification of statements and questions in each language, based on F0 cues.

6.2 Methods

6.2.1 Simulated Listeners

For each language, the exemplar-based model simulated twenty listeners: two per each of the ten listener orders in Table 5.4, corresponding to a male and a female human listener. The simulations of the male and female listeners in each listener order were different from each other because the trials were randomized in each phase for each human listener.

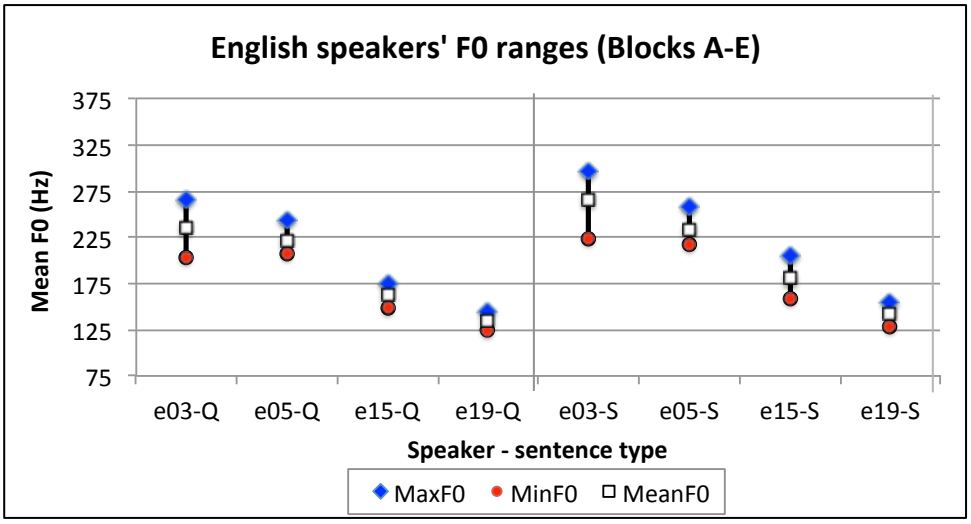
6.2.2 Stimuli

The model classified the same set of training and test stimuli used in the human listeners' identification task: 80 pairs of statements and questions produced by four speakers, gated in five forms (Section 5.2.2). Since I wanted to test whether the model could handle variation in the speaker's utterances (without normalizing F0 for each speaker's voice), I

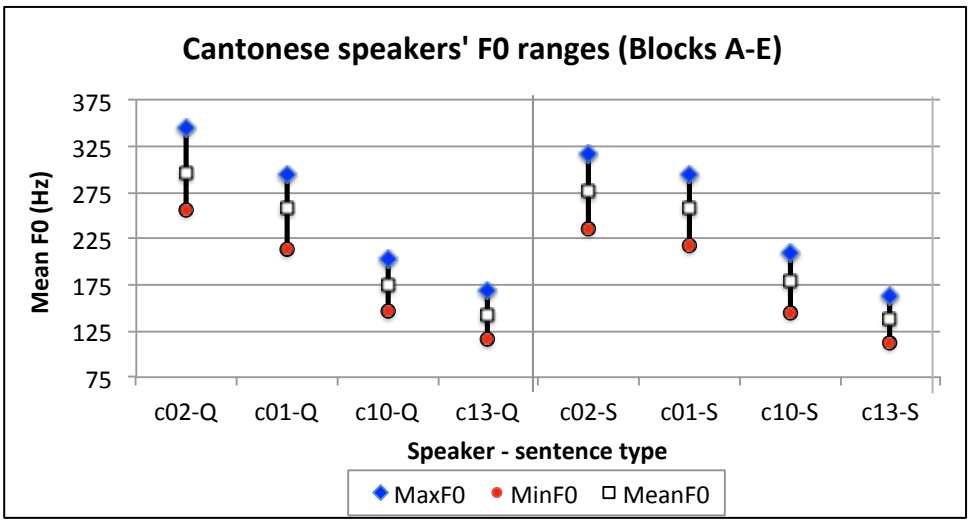
wanted to determine whether there was actually significant variation in the F0 ranges of the speakers I presented to the model. Figure 6.1 displays the F0 ranges of the four speakers' productions of the first two syllables in each language, averaged over blocks A to E. For the tone languages, the first two syllables contained both the high and low tones of the language (Table 4.3), which captured the speaker's initial F0 range for the utterance.

For each language, I ran six one-way ANOVAs with the maximum F0 (maxF0), minimum F0 (minF0), and mean F0 (meanF0) of each sentence type (statement, question) as the dependent measure, and with speaker as the independent factor. All of these ANOVAs revealed significant main effects of speaker (English: [$F(3, 76) > 71.7$, $p < 0.001$]; Cantonese: [$F(3, 76) > 59.8$, $p < 0.001$]; Mandarin: ($F(3, 76) > 46.7$, $p < 0.001$]). At $\alpha = 0.05$, post-hoc Tukey HSD tests revealed significant differences in the maxF0, minF0, and meanF0 of each sentence type among all except one pair of speakers from each language. Specifically, no significant difference was found 1) between the minF0 for English speakers e03 and e05's statements and questions, 2) between the minF0 and meanF0 for Cantonese speakers c01 and c02's statements, and 3) between the maxF0 for Mandarin speakers m07 and m14's statements. Overall, these results indicated that each language's speakers had relatively different F0 ranges among themselves and between the two sentence types.

(a)



(b)



(c)

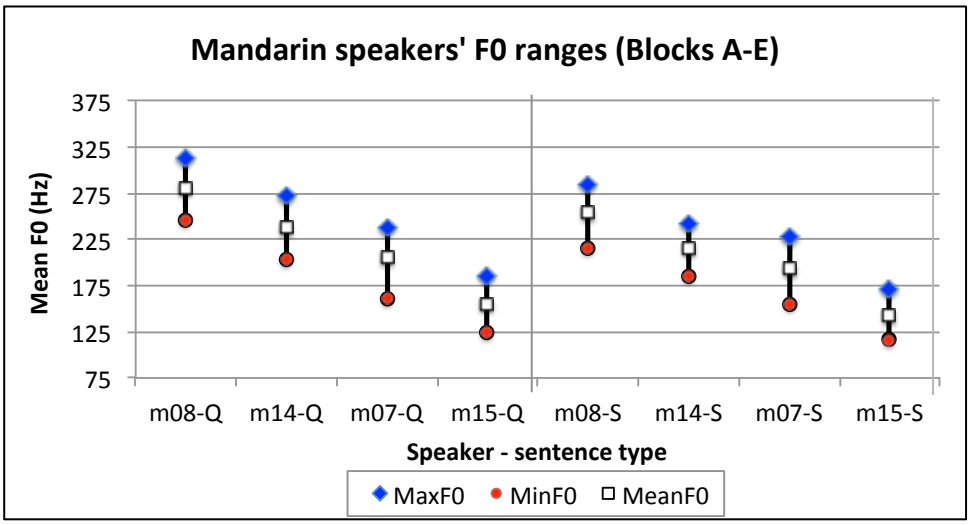


Figure 6.1. F0 ranges of the speakers' production of the first two syllables.

6.2.3 Classification Task

Similar to the training and testing conditions of the human listeners in the identification task, the training and testing of the computer model in the classification task used a two-fold cross-validation. The two runs of the cross-validation simulated sessions 1 and 2 of the perception study. In principle, the computer model performed the same training and testing tasks as the human listeners. In implementation, however, the processes for these two listener groups (computer and human) differed. The human listeners were trained once in each session on the whole utterances only; they were not trained on the fragment utterances later on in the session because they had already experienced the fragment utterances from the whole utterances. They were also tested on stimulus types NoLast and Last together in Part V, and on stimulus types First and Last2 together in Part VII (Table 5.3). However, the testing of computer model on each stimulus type was implemented as a separate process of the simulation. As shown in Table 6.1, each model simulation involved three processes (Sim-a, Sim-b, and Sim-c) in run #1, and five processes (Sim-a, Sim-b, Sim-c, Sim-d, and Sim-e) in run #2. First, Sim-a simulated the training and testing on the stimulus type Whole (Parts II and III). Then, in each subsequent process prior to testing, the computer model's 'memory' was refreshed with the trained utterances from Part II but in the same gated form as the test stimuli for that process (Parts II-b, II-c, II-d, and II-e). The reason for implementing the simulation in this way was to simplify the computer programming of the simulation. Table 6.1 lists the simulation process for each stimulus type. Parts II, III, V, and VII correspond to those parts in the identification task listed in Table 5.3.

Table 6.1. The stimulus type and number of trials presented in each simulation.

Run #	Simulation Process	Part	Phase	Number of Trials	Stimulus Type
1, 2	Sim-a	II	Training	80	Whole
1, 2	Sim-a	III	Testing	80	Whole
1, 2	Sim-b	II-b	Training	80	NoLast
1, 2	Sim-b	V	Testing	80	NoLast
1, 2	Sim-c	II-c	Training	80	Last
1, 2	Sim-c	V	Testing	80	Last
2	Sim-d	II-d	Training	80	Last2
2	Sim-d	VII	Testing	80	Last2
2	Sim-e	II-e	Training	80	First
2	Sim-e	VII	Testing	80	First

Overall, the model performed forty simulations: two for each of the ten listener orders, multiplied by two runs for each order. Table 6.2 lists the stimulus set used in each simulation, corresponding to the stimulus set used in each phase for each human listener (Table 5.4).

Table 6.2. Stimulus sets used for runs #1 and #2 of the 2-fold cross-validation.

Listener Order	Run #1		Run #2	
	Training Stimuli	Test Stimuli	Training Stimuli	Test Stimuli
1	1a	1b	1b	1a
2	1b	1a	1a	1b
3	3a	3b	3b	3a
4	3b	3a	3a	3b
5	5a	5b	5b	5a
6	5b	5a	5a	5b
7	7a	7b	7b	7a
8	7b	7a	7a	7b
9	9a	9b	9b	9a
10	9b	9a	9a	9b

6.2.4 Procedure

The approach to the model's training was based on the assumption that every token the listener had heard during the training phase in the identification task was stored in memory along with its correct category; the simulation of the training exercise replicated the same 'experience' for the model. For each simulation process of the computer model (e.g., Sim-b), the corresponding training stimuli (e.g., of type NoLast) were stored as categorized exemplars in the model's 'memory' prior to simulating the testing phase.[21]

During the simulation of each testing phase, the tokens presented to the model followed the same order as the trials presented to the corresponding human listener in the identification task. Each simulation compared each new token in the test set with every token in the training set (i.e., with every stored exemplar in the model). Once a test token had been categorized (or 'experienced'), it became part of the stored exemplars and was available for comparison with subsequent test tokens.

6.2.5 Statistical Analysis

First, I converted the classification scores from the simulation tests to d-prime and beta using the formulas in (5.1) and (5.2). Then I conducted a series of ANOVAs on the combined data from the simulations and perception tests to find out if there were any significant differences between the two listener types (computer and human) in perceptual sensitivity and response bias. The α-criterion was 0.05 for these ANOVAs. Since this chapter focuses on how well the computer model performed in comparison to

[21] Training could also be implemented with each token presented to the model sequentially and allowing the model to categorize each presented token by comparing it with the previously presented tokens first before labeling it with the correct category. Since the analysis of the performance on training is beyond the scope of this thesis, the current, simpler approach is preferred.

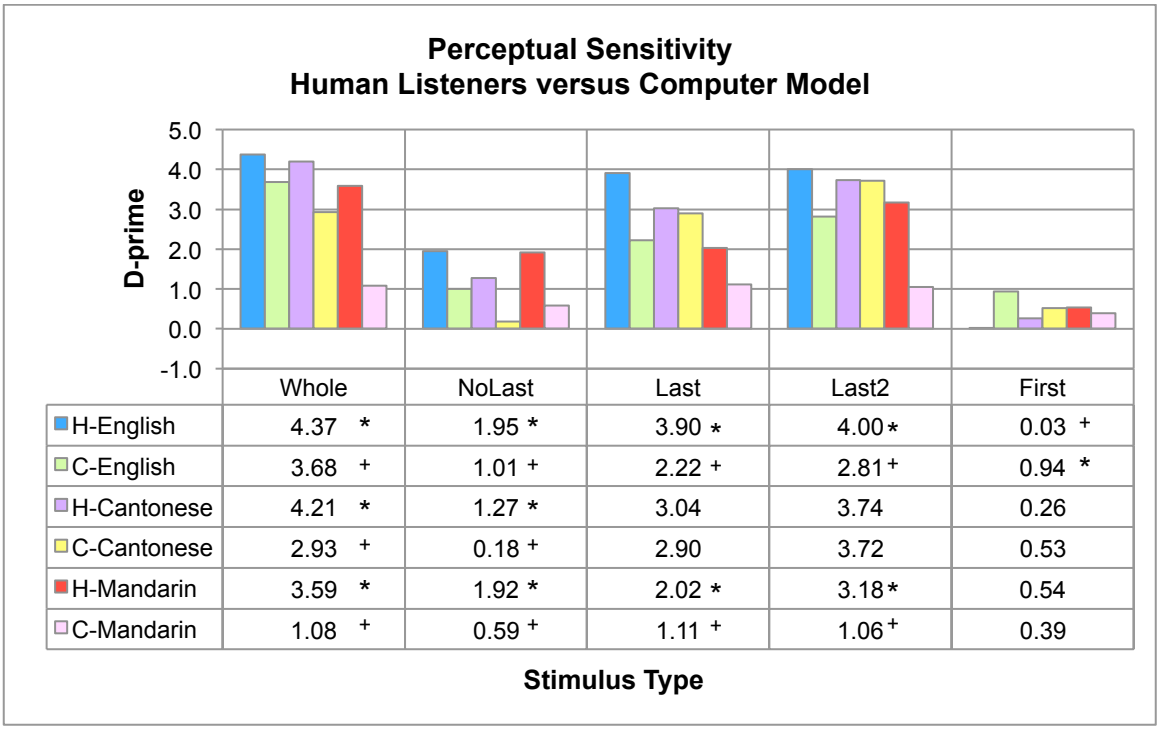
the human listeners, I only included significant effects and interactions involving the listener type factor in Section 6.3.

6.3 Results

6.3.1 Perceptual Sensitivity: Across Languages

To compare how well the computer model performed against the human listeners, I ran a three-way ANOVA with d' as the dependent measure, and with listener type, language, and stimulus type as independent factors. The results showed significant interactions between listener type and language [$F(2, 570) = 87.8$, $p < 0.001$], between listener type and stimulus type [$F(4, 570) = 107.7$, $p < 0.001$], and among listener type, language, and stimulus type [$F(8, 570) = 30.4$, $p < 0.001$].

A post-hoc Tukey HSD test revealed that, on all three languages, the human listeners ($\bar{x}$: English = 2.85, Cantonese = 2.50, and Mandarin = 2.25) performed significantly better than the computer model ($\bar{x}$: English = 2.13, Cantonese = 2.05, and Mandarin = 0.84). On four of the stimulus types, human listeners ($\bar{x}$: Whole = 4.06, NoLast = 1.72, Last = 2.99, and Last2 = 3.64) performed significantly better than the computer model ($\bar{x}$: Whole = 2.56, NoLast = 0.59, Last = 2.08, and Last2 = 2.53). However, on stimulus type First, the computer model ($\bar{x} = 0.62$) performed significantly better than the human listeners ($\bar{x} = 0.27$).



	Whole	NoLast	Last	Last2	First
H-English	4.37 *	1.95 *	3.90 *	4.00 *	0.03 +
C-English	3.68 +	1.01 +	2.22 +	2.81 +	0.94 *
H-Cantonese	4.21 *	1.27 *	3.04	3.74	0.26
C-Cantonese	2.93 +	0.18 +	2.90	3.72	0.53
H-Mandarin	3.59 *	1.92 *	2.02 *	3.18 *	0.54
C-Mandarin	1.08 +	0.59 +	1.11 +	1.06 +	0.39

Figure 6.2. Interaction among listener type, language, and stimulus type on d'.

Figure 6.2 displays the interaction among listener type, language, and stimulus type on d'. All three language listeners performed significantly better than the model on stimulus types Whole and NoLast. Additionally, the English and Mandarin listeners performed significantly better than the model on stimulus types Last and Last2. However, the model performed significantly better than the English listeners on stimulus type First.

The Tukey HSD test also revealed that, overall, the model performed significantly better on English and Cantonese than on Mandarin. Its performance on the stimulus types, from significantly best to worst, was as follows: Whole, Last2 > Last > NoLast, First. Figure 6.3 displays the interaction between language and stimulus type on d' for the model. The model performed significantly better on both English and Cantonese than on Mandarin in classifying stimuli that included the final syllable (Whole, Last, and Last2). It also performed significantly better on English than Cantonese in classifying stimuli of

types Whole and NoLast, but worse in classifying stimuli of types Last and Last2. In addition, it performed significantly better on English than Mandarin in classifying stimuli of type First.

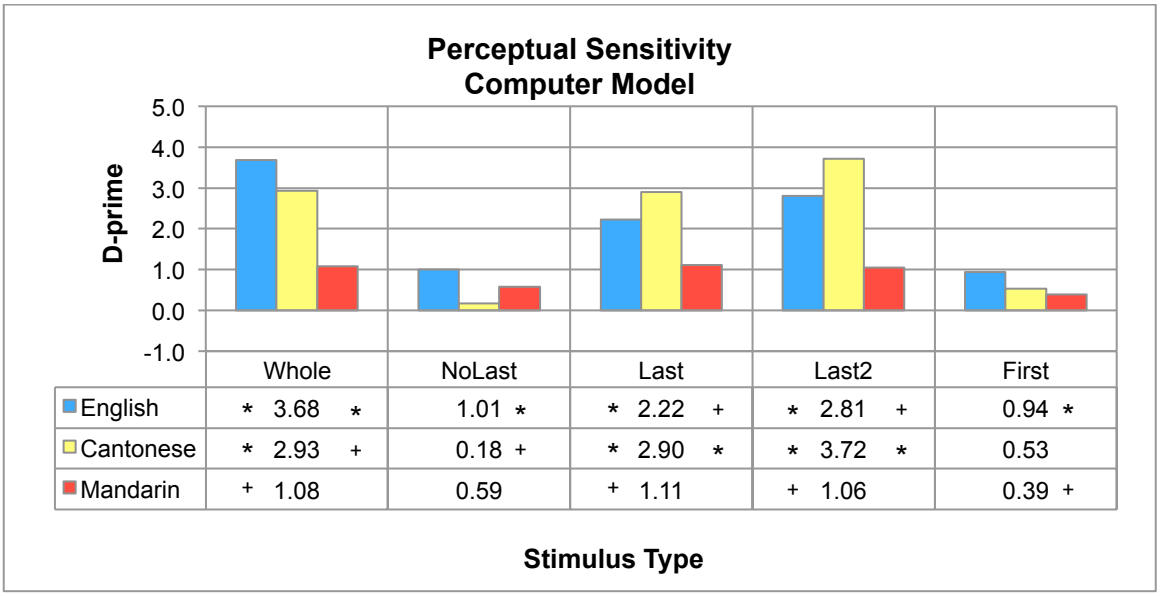


Figure 6.3. Interaction between language and stimulus type on d', model only.

In general, the model performed significantly better on stimulus types that included the final syllable (Whole, Last2, and Last) than on stimulus types that excluded it (NoLast and First), as shown in Table 6.3.

Table 6.3. Interaction between language and stimulus type on d', model only.

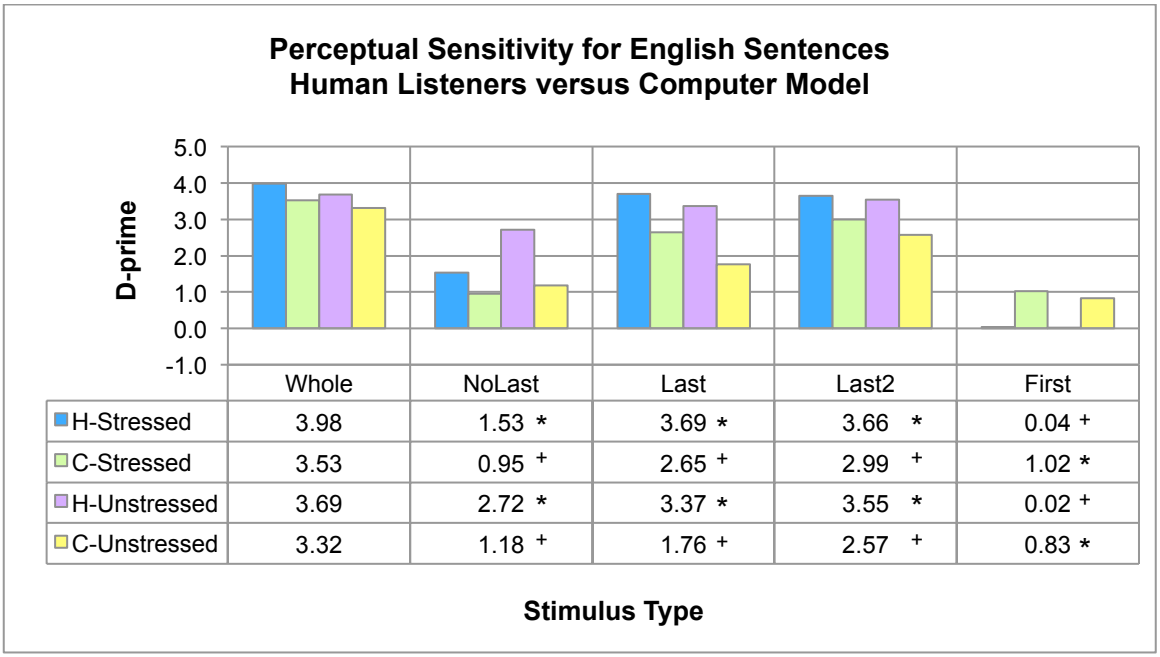
Language	**Stimulus Types** (ordered by mean d' scores)
English	Whole > Last2 > Last > NoLast, First
Cantonese	Last2 > Whole, Last > NoLast, First
Mandarin	Whole, Last2, Last > NoLast, First

Among the stimuli that included the final syllable, the model performed significantly better on the longer than shorter stimuli in English and on the final two syllables alone than any other stimuli in Cantonese, but performed equally on all three stimulus types in Mandarin.

6.3.2 Perceptual Sensitivity: Effect of Final Stress/Tone

In order to compare how the suprasegmental structures of stress and tone affected the computer model's performance, I ran a three-way ANOVA for each language, with d' as the dependent measure and with listener type, stimulus type, and stress/tone as independent factors. Significant interactions among listener type, stimulus type, and stress/tone were found in English [$F(4, 380) = 4.9$, $p < 0.001$], Cantonese [$F(12, 760) = 2.9$, $p < 0.001$], and Mandarin [$F(12, 760) = 3.7$, $p < 0.001$].

Post-hoc Tukey HSD tests revealed the following significant results, as indicated in Figures 6.4-6.9. For English (Figure 6.4), the human listeners performed significantly better than the computer model on stimulus types NoLast, Last, and Last2, but worse on stimulus type First, regardless of whether the sentence-final syllable was stressed or unstressed. There was no significant difference between the English listeners and the computer model on stimulus type Whole.



	Whole	NoLast	Last	Last2	First
H-Stressed	3.98	1.53 *	3.69 *	3.66 *	0.04 +
C-Stressed	3.53	0.95 +	2.65 +	2.99 +	1.02 *
H-Unstressed	3.69	2.72 *	3.37 *	3.55 *	0.02 +
C-Unstressed	3.32	1.18 +	1.76 +	2.57 +	0.83 *

Figure 6.4. Interaction among listener type, stimulus type, and stress on d' for English.

In addition, as shown in Figure 6.5, the computer model performed significantly better on stimulus type Last when the syllable was stressed than when it was unstressed.

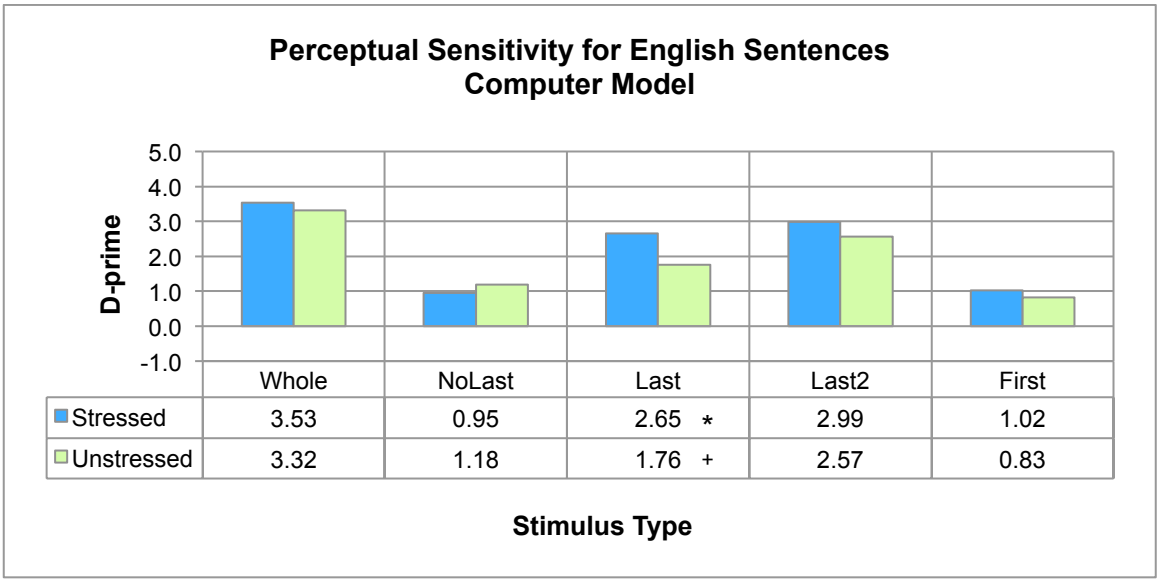
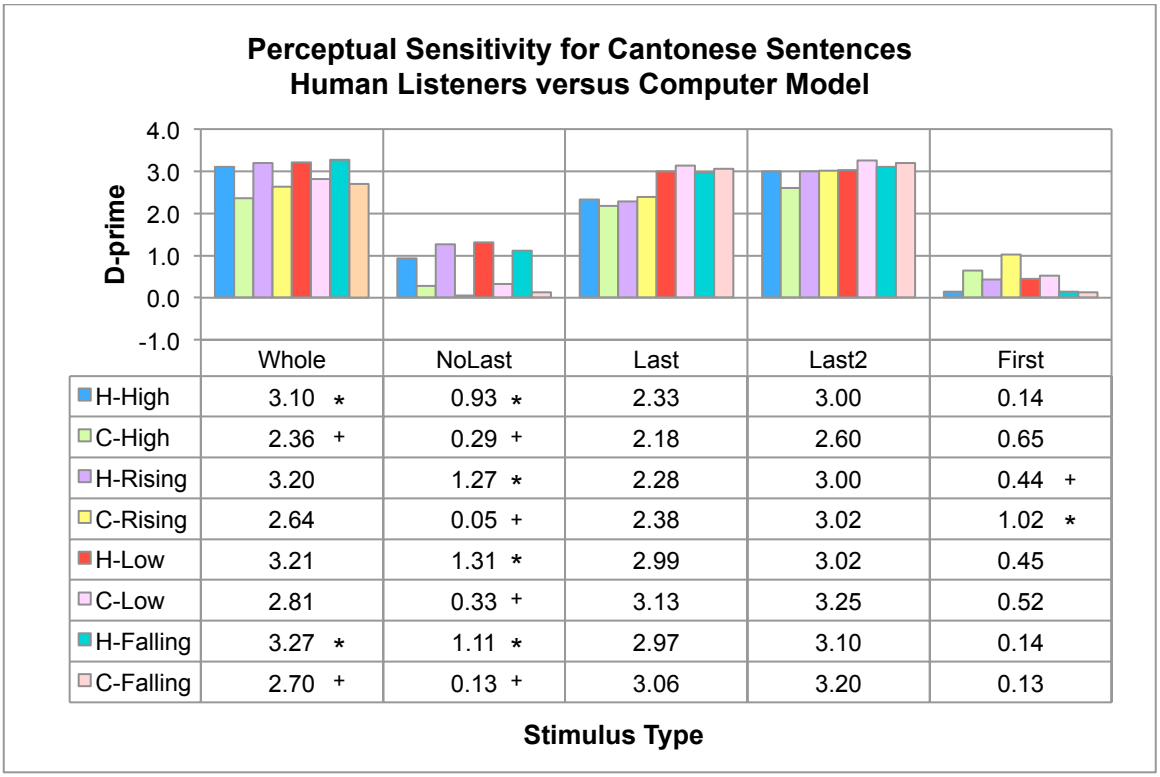


Figure 6.5. Interaction between stimulus type and stress on d' for English, model only.

For Cantonese (Figure 6.6), the human listeners performed significantly better than the computer model on stimulus type NoLast, regardless of the tone on the missing syllable. They also performed significantly better than the model on stimulus type Whole when the final syllable carried a high or falling tone. The model performed significantly better than the Cantonese listeners only on stimulus type First when the missing final syllable carried a rising tone.[22]

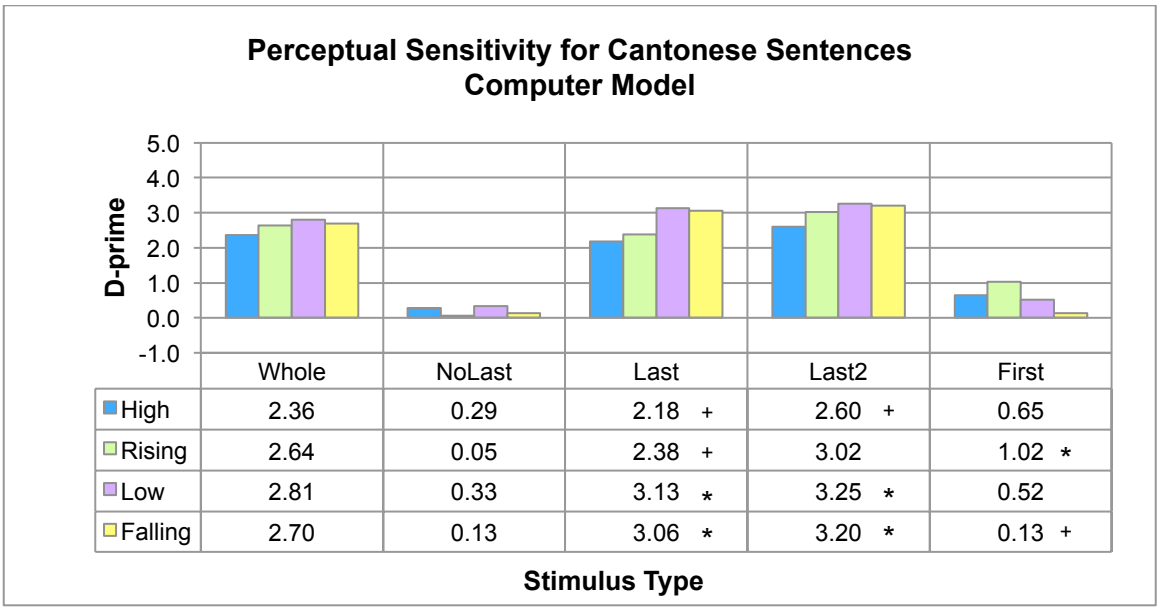


	Whole	NoLast	Last	Last2	First
H-High	3.10 *	0.93 *	2.33	3.00	0.14
C-High	2.36 +	0.29 +	2.18	2.60	0.65
H-Rising	3.20	1.27 *	2.28	3.00	0.44 +
C-Rising	2.64	0.05 +	2.38	3.02	1.02 *
H-Low	3.21	1.31 *	2.99	3.02	0.45
C-Low	2.81	0.33 +	3.13	3.25	0.52
H-Falling	3.27 *	1.11 *	2.97	3.10	0.14
C-Falling	2.70 +	0.13 +	3.06	3.20	0.13

Figure 6.6. Interaction among listener type, stimulus type, and tone on d' for Cantonese.

[22] For the stimulus type First in Cantonese, the F0 difference between statements and questions on the second syllable is greater when the sentence-final tone is rising rather than high, low, or falling (figure not shown). This effect is most likely due to the coarticulation of the tones between the second syllable of stimulus type First and the syllable that follows it. Although the initial two syllables of all of the target sentences within a dialogue are the same regardless of the final tone of these sentences, the third syllable could be different among these sentences. The computer model does not distinguish between tonal effects and intonational effects, but the human listeners could and would likely ignore any tonal effects that are irrelevant to the identification of the sentence intonation.

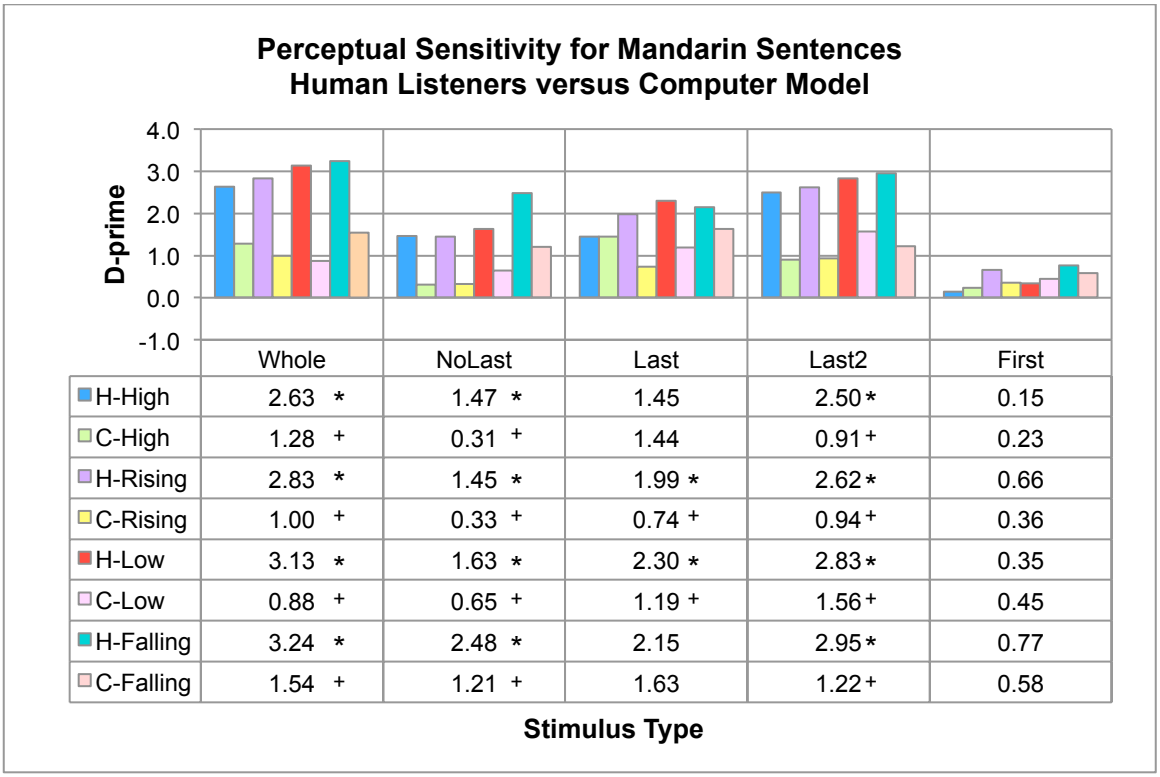
In addition, as shown in Figure 6.7, the computer model performed significantly better on stimulus type Last when this syllable carried a low or falling tone, rather than a high or rising tone, and on stimulus type Last2 when the final syllable carried a low or falling tone, rather than a high tone. The model also performed significantly better on stimulus type First when the missing final syllable carried a rising tone, as opposed to a falling tone.



	Whole	NoLast	Last	Last2	First
High	2.36	0.29	2.18 +	2.60 +	0.65
Rising	2.64	0.05	2.38 +	3.02	1.02 *
Low	2.81	0.33	3.13 *	3.25 *	0.52
Falling	2.70	0.13	3.06 *	3.20 *	0.13 +

Figure 6.7. Interaction between stimulus type and tone on d' for Cantonese, model only.

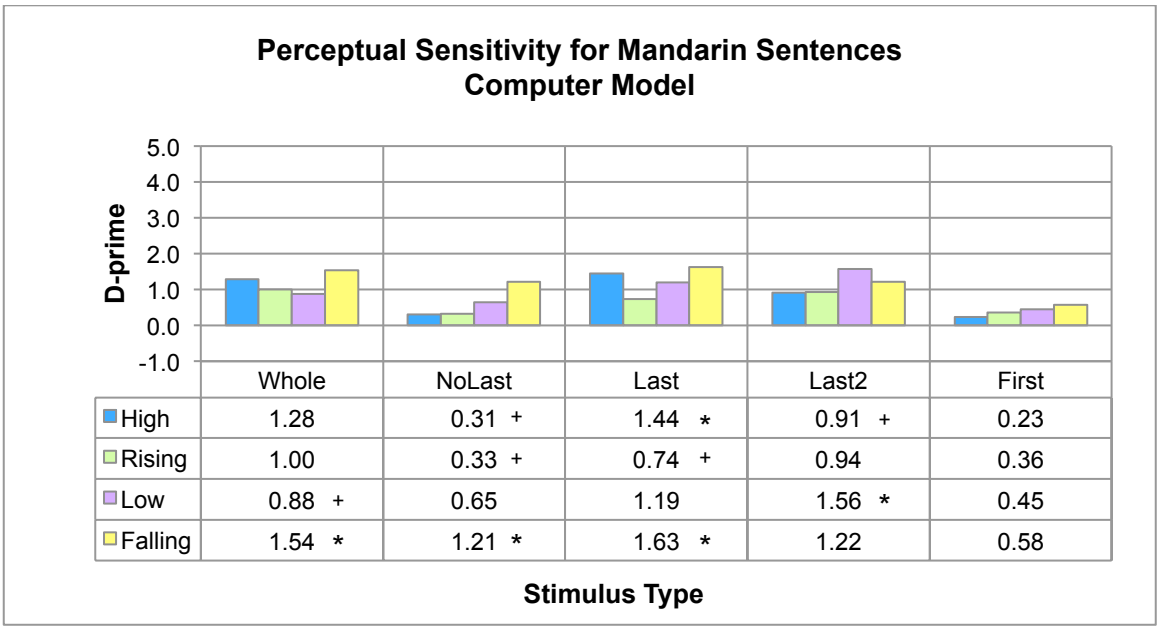
For Mandarin (Figure 6.8), the human listeners performed significantly better than the computer model on stimulus types Whole, NoLast, and Last2, regardless of the tone on the final syllable. They also performed significantly better than the model on stimulus type Last when the syllable carried a rising or low tone. There was no significant difference between human performance and computer performance on stimulus type First.



	Whole	NoLast	Last	Last2	First
H-High	2.63 *	1.47 *	1.45	2.50 *	0.15
C-High	1.28 +	0.31 +	1.44	0.91 +	0.23
H-Rising	2.83 *	1.45 *	1.99 *	2.62 *	0.66
C-Rising	1.00 +	0.33 +	0.74 +	0.94 +	0.36
H-Low	3.13 *	1.63 *	2.30 *	2.83 *	0.35
C-Low	0.88 +	0.65 +	1.19 +	1.56 +	0.45
H-Falling	3.24 *	2.48 *	2.15	2.95 *	0.77
C-Falling	1.54 +	1.21 +	1.63	1.22 +	0.58

Figure 6.8. Interaction among listener type, stimulus type, and tone on d' for Mandarin.

In addition, as shown in Figure 6.9, the computer model performed significantly better on stimulus type Whole when the final syllable carried a falling tone rather than a low tone, on stimulus type Last when the syllable carried a high or falling tone rather than a rising tone, and on stimulus type Last2 when the final syllable carried a low tone rather than a high tone. The model also performed significantly better on stimulus type NoLast when the missing final syllable had a falling tone rather than a high or rising tone.



	Whole	NoLast	Last	Last2	First
High	1.28	0.31 +	1.44 *	0.91 +	0.23
Rising	1.00	0.33 +	0.74 +	0.94	0.36
Low	0.88 +	0.65	1.19	1.56 *	0.45
Falling	1.54 *	1.21 *	1.63 *	1.22	0.58

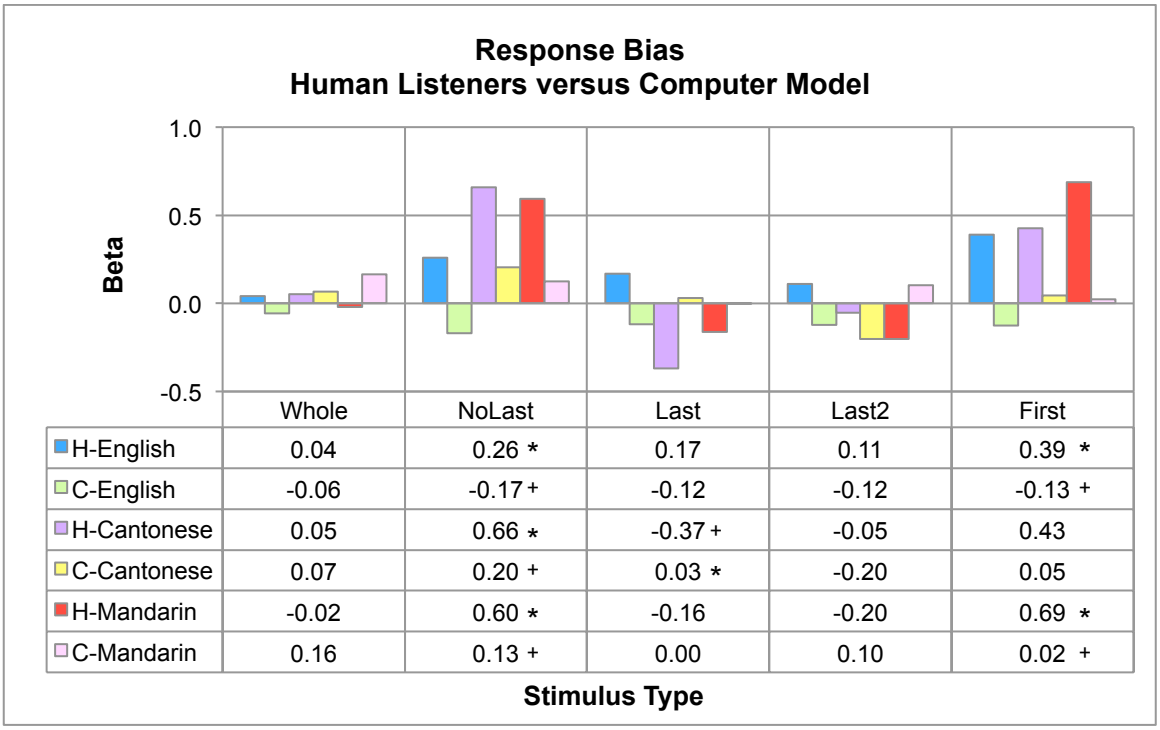
Figure 6.9. Interaction between stimulus type and tone on d' for Mandarin, model only.

6.3.3 Response Bias: Across Languages

In order to compare the human listeners' and the computer model's response bias, I ran a three-way ANOVA with ß as a dependent measure and with listener type, language, and stimulus type as independent factors. The results showed a significant main effect of listener type [$F(1, 570) = 41.3$, $p < 0.001$]. There were also significant interactions between listener type and language [$F(2, 570) = 6.5$, $p < 0.01$], between listener type and stimulus type [$F(4, 570) = 22.4$, $p < 0.001$], and among listener type, language, and stimulus type [$F(8, 570) = 3.9$, $p < 0.001$].

A post-hoc Tukey HSD test revealed that, overall, the human listeners ($\bar{x} = 0.17$) were significantly more biased towards statements than questions, compared to the computer model ($\bar{x} = 0.00$). Among the three languages, the human listeners ($\bar{x} = 0.19$) were significantly more biased towards statements than questions in English, compared to the computer model ($\bar{x} = -0.12$). There was no significant difference in ß between the Cantonese and Mandarin listeners and the computer model. Among the different stimulus types, the human listeners were significantly more biased towards statements than questions, compared to the computer model, on stimulus types NoLast ($\bar{x}$: humans = 0.50 vs. computer = 0.05) and First ($\bar{x}$: humans = 0.50 vs. computer = -0.02).

Figure 6.10 displays the interaction among listener type, language, and stimulus type on ß. The English and Mandarin listeners were significantly more biased towards statements than questions, compared to the computer model, on stimulus type NoLast and First. The Cantonese listeners were significantly more biased towards statements than questions, compared to the computer model, on stimulus type NoLast only. However, the computer model showed more response bias towards statements than the Cantonese listeners did on stimulus type Last.



	Whole	NoLast	Last	Last2	First
H-English	0.04	0.26 *	0.17	0.11	0.39 *
C-English	-0.06	-0.17 +	-0.12	-0.12	-0.13 +
H-Cantonese	0.05	0.66 *	-0.37 +	-0.05	0.43
C-Cantonese	0.07	0.20 +	0.03 *	-0.20	0.05
H-Mandarin	-0.02	0.60 *	-0.16	-0.20	0.69 *
C-Mandarin	0.16	0.13 +	0.00	0.10	0.02 +

Figure 6.10. Interaction among listener type, language, and stimulus type on ß. Overall, the computer model responded to the Cantonese and Mandarin stimuli with significantly more bias towards statements than questions than it did to the English stimuli. The model showed no significant difference in response bias among the three languages by stimulus type.

6.4 Discussion

6.4.1 The Computer Model's Performance

The computer model accurately classified the sentence type of statements and questions well above chance in English and Cantonese. Since the model classified these sentence types using only a time series of F0 values extracted from the utterance's intonation contour (the MpC method discussed in Chapter 4), this result suggests that F0 height and direction contribute to the identification of statements and questions in these languages. Although the relative duration of the significant F0 differences between statements and questions is longer in Mandarin than in English and Cantonese overall (Figure 5.2), the model only performed slightly above chance in Mandarin, possibly because the longer the cue, the less salient it is to the model. The current model is not formulated to deal with time variation (see Chapter 7 for suggested enhancements to the model).

The computer model demonstrated that it was indeed sensitive to the question cues in the final syllable in all three languages because it was more accurate in classifying stimuli that included the final syllable than stimuli that excluded it. The model was also sensitive to the language-specific F0 cues for questions in these languages even though it had not been given an abstract representation of their intonation patterns. First of all, the model performed significantly better on the final two syllables combined than on the final syllable alone in English, suggesting that it was sensitive to the timing difference of the final stress: the final stress occurs 100% of the time in the final two syllables combined but only part of the time in the final syllable alone. Secondly, in Cantonese, the model performed best on the final two syllables combined, suggesting that it was sensitive not only to the high F0 rise in the final syllable but also the F0 context in

the preceding syllable. It did not perform as well on whole utterances as the final two syllables, however, possibly because the statement and question contours overlapped each other in the first 40% of the utterances from time points 1 to 5 (Figure 5.2). Since the model used all F0 values at the eleven time points in determining a token's similarity to the statement and question categories, this non-distinctive stretch of the utterance might have reduced the relative contribution of the F0 cue in the final two syllables. Thirdly, in Mandarin, the model performed equally well among all three types of stimuli that included the final syllable regardless of their lengths, suggesting that the model was sensitive to Mandarin's raised pitch range cue throughout the utterance. It also performed equally well between the two types of stimuli that excluded the final syllable in Mandarin, likely due to the raised pitch range cue as well. Finally, the model performed the same on both types of non-final stimuli in English and Cantonese, suggesting that it was also sensitive to the absence of a salient F0 cue for distinguishing between statements and questions in these languages, prior to the penultimate syllables.

6.4.2 The Computer Model versus Human Listeners

In all three languages, the human listeners performed significantly better than the computer model. Since these listeners were presented with more acoustic information in their stimuli than just the eleven F0 values presented to the model, this result suggests that, in addition to F0, the human listeners were using other acoustic cues to differentiate questions from statements (e.g., talking speed or duration of the final syllable). For example, even though both the computer model and the human listeners performed significantly better on stimuli that included the final syllable than on stimuli that

excluded it, the human listeners also performed better on the longer stimuli than the shorter stimuli in each of these final and non-final stimulus groups. As mentioned in Section 6.4.1, the mid-portion of the utterance did not seem to have improved the model's performance in classifying the stimuli.

More specifically, the human listeners performed significantly better than the computer model on all stimuli, except for the final syllable and final two syllables in Cantonese, as well as the first two syllables in all three languages. Presumably, the human listeners used their prior experience in their native prosodic systems—which the model lacked—to help them identify the sentence types of these stimuli. The computer model, however, was able to perform nearly as well as the human listeners on the final (two) syllables in Cantonese, most likely because of its high F0 rise at the end of the utterance for questions. In other words, the model was able to perform nearly as well as the human listeners on these stimuli based on the significant F0 differences between the statements and questions, while not having to deal with the timing variation of the English question cue and the tonal variation of the Mandarin question cue. In addition, the computer model did not perform significantly differently than the human listeners on the first two syllables in Cantonese and Mandarin. For the human listeners, there might not have been other acoustic cues available to the statement/question distinction for these stimuli or it could have been difficult to match up these fragments of utterances with their source utterances in memory.

In addition, there were specific parts of the utterance where the human listeners performed worse than the computer model, for example, on the first two syllables in English. These syllables had higher F0s in statements than questions (Figure 5.3). As

described in Section 2.6.2, English questions end with a falling intonation (H* H- L%), whereas statements end with a rising intonation (L* L- H%). Perhaps in anticipating the final fall or rise of the utterance, the speakers on average produced these sentences initially with a higher pitch or lower pitch, respectively. Since speakers of many languages tend to associate higher F0 with questions and lower F0 with statements, these reversed F0 patterns might have confused the native English listeners, who were familiar with questions having higher F0. The model, however, had no such prior experience with the English intonation and performed the classification task based on just the F0 cues given. This analysis is also supported by the fact that the Mandarin listeners performed significantly better than the English listeners on the first two syllables, which had higher F0 in questions than statements in Mandarin. The model performed significantly better on English than Mandarin on these two syllables, possibly due to tonal variation in the Mandarin stimuli. Similarly, the English listeners performed significantly better than the Cantonese listeners on the final syllable alone, likely because the Cantonese listeners confused the rising tones on statements with question rises. The model, however, performed significantly better on Cantonese than English on the final two syllables and the final syllable alone, likely because the timing of the question cue is more consistent in Cantonese than in English. In summary, native experience of the prosodic systems might have improved the human listeners' performance in the identification task in general but might have also worsened it in certain cases.

6.4.3 Cross-linguistic Performance

Cross-linguistically, the computer model performed significantly better on English and Cantonese than on Mandarin. Based on the human listeners' d-prime scores in the identification task, however, the English listeners performed significantly the best, followed by the Cantonese listeners, and then the Mandarin listeners. The model's considerably lower d-prime of 0.84 for Mandarin, compared to its d-prime of 2.13 and 2.05 for English and Cantonese, respectively, indicates that the F0 patterns in signaling statements and questions are meaningfully different for Mandarin than for English and Cantonese. The slightly lower d-prime of 2.25 for the Mandarin listeners, compared to the d-prime of 2.85 and 2.50 for the English and Cantonese listeners, respectively, suggests that the F0 cues for questions are slightly more difficult for native listeners to detect in Mandarin than in English and Cantonese. In other words, Mandarin's raised pitch range cue is more difficult for native listeners to detect than English's and Cantonese's final high F0 rises. The fact that the model was sensitive to the subtle F0 differences between statements and questions in the first two syllables of the English and Mandarin stimuli but performed better in English (d' = 0.94) than in Mandarin (d' = 0.39) suggests that other factors, such as the effect of tones on intonation, might have made the Mandarin question cue harder to detect. Likewise, the slightly lower d-prime of 2.50 for the Cantonese listeners than the d-prime of 2.85 for the English listeners also suggests that there was a similar tonal effect on the detection of the Cantonese question cue. However, this effect did not significantly impact the model's performance on these two languages, likely due to the timing of their question cues (as was mentioned in Section 6.4.2). Both the computer model and the human listeners also performed significantly

better on stimuli that excluded just the final syllable in English than in Cantonese, most likely due to the timing of the Cantonese question rise at the end of the utterance. Thus, the model's simulation results, analyzed alongside the human listeners' results, provided strong evidence that there are meaningful differences in how human listeners perceive question intonation in different languages.

6.4.4 Effects of Final Stress/Tone on Intonation Perception

In English, the computer model performed significantly better on the final syllable when it was stressed than unstressed, whereas the human listeners performed significantly better on stimuli that excluded the final syllable when the excluded syllable was unstressed rather than stressed. When the final syllable was stressed, the entire question rise occurred in the final syllable. When the final syllable was unstressed, the initial part of the question rise occurred in the penultimate syllable, and the final part of the question rise occurred in the final syllable. The model and the English listeners' performances, then, suggest that the model required the entire question rise in the signal to identify the sentence types of the stimuli, whereas the human listeners could perform equally well with either just part of the question rise or the entire question rise in the signal. Native experience likely helped the human listeners to recover part of the missing cue.

In Cantonese, the significantly better performance of both the computer model and the human listeners on final syllables that carried a low or falling tone as opposed to a high or rising tone suggests that a final low or falling tone was easier to interpret than a final high or rising tone. Interestingly, the computer model also performed significantly better on the Last2 stimuli that carried a final low or falling tone as opposed to a final

high tone, but the Cantonese listeners did not. Since the F0 or tonal context of the final tone appeared to have helped both the computer model (Section 6.4.1) and the Cantonese listeners (Section 5.4.3) in identifying the sentence types of the Last2 stimuli, this result suggests that other contextual information, such as the semantics of the penultimate syllable might have helped the human listeners. For example, both the final syllable *si4* 'time' in the question shown in Figure 2.7 and the final syllable *si2* 'chronicle' in the statement shown in Figure 2.8 have similarly rising contours, which makes it difficult to identify their sentence types. However, when the preceding syllable is included in the stimuli, it is possible for the Cantonese listeners to determine the final tone from the semantics of combined expressions (i.e., *zeon2 si4* 'on time' and *lik6 si2* 'history'). Knowing the target tone could in turn help to match the F0 contour with the sentence type; for example, the canonical tone of *si4* 'time' is falling, so if the F0 contour of *zeon2 si4* 'on time' rises at the end, this stimulus must be a question.

In addition, the Cantonese listeners performed significantly better than the computer model on whole utterances that ended in a high or falling tone. On the one hand, the computer model performed the worst on whole utterances that ended in a high tone (Figure 6.6), most likely because the statement and question contours overlapped each other through more of these utterances than they did in any of the other utterances (figure not shown). On the other hand, the Cantonese listeners performed the best on whole utterances that ended in a falling tone, most likely because the statement intonation of these utterances ended in a fall while the question intonation ended in a rise (Figure 5.4). The falling tone (Tone 4) is the only Cantonese tone that is consistently realized as falling on the final syllable of a statement.

In Mandarin, the tones retain their canonical forms in questions, so they are less effective (salient) cues for this task than the overall F0 difference between statements and questions. Even so, the human listeners performed significantly better than the computer model on the final syllable when the syllable carried a low or rising tone. The similarity of the falling and rising contours of these tones on questions (Figure 5.4) might have confused the computer model (d' = 0.74 for T3; d' = 1.19 for T4) (Figure 6.8). It is likely that experience with the combination of these native tones and intonation, as well as the human ability to perceive the timing cues, helped the Mandarin listeners identify these sentence types (d' = 2.30 for T3; d' = 2.15 for T4). For example, low tones are realized as falling-rising [214] at the end of questions and as falling [21] at the end of statements; as Figure 5.4 shows, the timing of the dip in the question contour for [214] is later than the dip in the question contour for the rising tone. Generally, both the computer model and the human listeners performed significantly better on final syllables that carried a falling tone, rather than a high or rising tone and on stimuli that excluded final syllables that carried a falling tone than a high or rising tone. As discussed in Section 5.4.3, the onset of the falling tone is raised much higher relative to the rest of the falling tone in questions than in statements, enabling it to be detected much easier than the other tones. Thus, some acoustic differences can be detected without native experience, such as the greater F0 difference between statements and questions (e.g., the Mandarin falling tone), whereas others (e.g., the realization of the Mandarin low tone as [214] question-finally) require native sensitivity, which increases with more relevant language-specific knowledge and experience.

6.4.5 Listeners' Response Bias

For stimuli that excluded the final syllable, the English and Mandarin listeners showed significantly more bias towards statements than questions, compared to the computer model. Since the question rise occurs mainly in the final syllable, this result suggests that the statement is the unmarked response type for these listeners, possibly because speakers generally encounter more statements than questions outside of laboratory settings. The model, however, showed only a slight bias for these stimuli in Mandarin but significantly more bias towards questions than statements for these stimuli in English (Figure 6.10). Since the model has no prior linguistic 'experience' other than the stimuli that were previously presented to it during training and testing, its biases are shaped by the calculation of the similarity between the statement and question contours, specifically, their F0 values. Since the first two syllables of the English utterances had higher F0 in statements than questions—but not the rest of the utterance—this F0 pattern likely influenced the model's bias towards questions for these stimuli. In Mandarin, however, the questions were clearly higher in F0 than statements, so the model was less likely to have developed strong biases from these F0 contours.

In Cantonese, the human listeners showed significantly more bias towards statements than questions for stimuli that excluded just the final syllable, but more bias towards questions than statements for the final syllable alone, compared to the computer model. The difference in response bias between the Cantonese listeners (ß = -0.37) and the computer model (ß = 0.03) on the final syllable suggests that these human listeners might have perceived the question cue in the final syllable differently, as opposed to the model. The fact that the model performed with nearly no bias suggests that there are

discernible differences between the statement rises and question rises in the final syllable. However, the Cantonese listeners' rich experience with the questions in their native language over time might have led them to develop a strong association between questions and a final F0 rise. This mental association, in turn, might have affected their performance behaviour in the identification task, such that they might have paid less attention to the actual F0 signal in the final syllables. This bias might have also led these listeners to respond with more bias towards statements than questions on stimuli that excluded just the final syllable due to the lack of a final F0 rise. Thus, the categorization of an utterance is based on more than just the acoustic signal alone; the listener's bias is also involved in the activation of the exemplars in memory (Johnson, 1997). The fact that the English and Mandarin listeners did not exhibit a significant bias towards question responses on the final syllable alone, compared to the computer model, suggests that this particular bias is language-specific.

6.4.6 The Human Listeners' Perception of Intonation

As I expected, the human listeners performed significantly better than the computer model on the same identification task in determining the sentence type of statements and questions, based solely on intonation. Importantly, the differences in performance on this task between the human listeners and the computer model (as discussed earlier in this Section 6.4) reveal meaningful information on how these human, native listeners process intonation in statements and questions. First of all, the English listeners were able to deal with the timing differences of the question cues between utterances in English, which means that these human listeners were able to align the utterances based on the salient

intonation cues of these utterances, whether a final fall or a final rise. They would probably need to perform this time alignment before comparing two utterances to determine how similar they are with each other. Secondly, the human listeners were able to identify the sentence type of an utterance fragment with only partial cue (e.g., the initial part of the question rise) in the intonation, which means that these listeners had stored the whole utterances that they had experienced during training in memory and then accessed them to match with the utterance fragments in order to determine the sentence types of these fragments. Thirdly, the human listeners used contextual information (e.g., the phonetic information of the adjacent syllable) to help determine the sentence-type cue of an utterance, which again suggests that they stored these utterances in their original, whole form. It also suggests that the stored utterances were detailed and that the listeners would use all available information from the stored utterances to process an utterance that they had just heard. Fourthly, the human listeners were paying less attention to portions of the utterance where the sentence-type cues were less salient (e.g., the mid-portion of an utterance), which suggests that although they would store experienced utterances in fine acoustic detail, they would focus on relevant information only when processing a new utterance. Fifthly, both the Cantonese and Mandarin listeners performed differently in identifying the sentence types of utterances, across all final tones. This difference in performance suggests that a tone-general intonation pattern for the sentence types (e.g., a final F0 rise or H% for Cantonese questions) was inadequate and that these listeners were using both the knowledge of their native tonal systems and the detailed contours of their experienced tone-specific utterances to help them identify the sentence intonation; in other words, the specific variability is perceptually useful. Lastly, the human listeners

showed language-specific biases in responding to the stimuli in the identification task (e.g., the Cantonese listeners were more biased towards questions that statements when responding to the final-syllable stimuli). This response behaviour suggests that the perception of intonation is affected by listeners' bias and that this bias could decrease (in the case of the Cantonese listeners hearing final syllables) or increase the listeners' attention on (certain parts of) the acoustic signal of the utterance.

6.5 Summary

This chapter has examined how well an exemplar-based computational model performed in classifying statements and questions in English, Cantonese, and Mandarin, compared to the human listeners. Even though the proposed exemplar-based model is simple, it successfully classified statements and questions 1) without F0 normalization of the speakers' voices, 2) without knowing explicitly where to look for statement and question intonation cues, 3) without knowledge of the stress and tonal patterns that influenced the surface intonation in each utterance, 4) using only eleven F0 values of the intonation contour, and 5) using the same set of auditory properties and the same classification method for all three different intonation systems. Overall, the human listeners did perform significantly better than the model, suggesting that experience with the native language might have helped the listeners perform the identification task, and that these listeners might be processing the intonational information in the stimuli differently than the computer model did. The next chapter provides some suggestions for improving the model.

Chapter 7: Towards a Generalized Intonation Perception Model

7.1 The Kernel Model

In extending exemplar theory to intonation perception, I designed a computational model that categorizes statements and echo questions based on F0 alone, using a simplified version of the algorithm adapted from Nosofsky (1988) and Johnson (1997). This simple design enabled the model to achieve its three core purposes: 1) to simulate intonation perception, 2) to accommodate multiple languages, and 3) to model human performance as a means of improving our scientific understanding of how human listeners categorize different sentence types through intonation.

To enable the model to classify statements and questions based on intonation, I used the main linguistic correlate of intonation, F0, in the similarity calculation to determine category membership. To account for the rise and fall in pitch in intonation in the similarity calculation, I extracted F0 measurements at eleven equidistant time points of the intonation contour, capturing its F0 shape quantitatively. To enable the model to perform well overall on all three languages, the model also used only auditory properties that were common to the intonation cues of all three languages: the differences in F0 between the statement and question intonation contours. However, human listeners have access to a wealth of acoustic and non-acoustic information in speech above and beyond F0 (e.g., speaker identity, gender, and duration). To enable the model to approach the performance of the human listeners in the intonation perception task, it would be necessary to include this information in the similarity calculation. However, I opted for a bare model, a kernel model, to find out how well the model does without tweaking the similarity calculation to account for these factors. A benefit of this implementation is that

the performance of the kernel model can be used as a benchmark for more advanced models.

Chapter 6 demonstrated that this exemplar-based model was successful in classifying the statements and echo questions for all three languages. Although this kernel model did not perform as well as the human listeners in the classification task, it mirrored the performance patterns of the human listeners in many aspects (Section 6.4). The human listeners' abilities to succeed in the following also account for why they performed considerably better than the computer model in the identification task:

1) they can deal with the different speaking rates,
2) they can deal with the timing differences of the question cues between utterances,
3) they can recognize partial question cues,
4) they can use contextual information to help identify the question cue,
5) they can handle question cues that extend throughout an entire utterance,
6) they can ignore parts of the utterance that contribute little to the intonation distinction,
7) they can use other acoustic cues, such as duration of the final syllable, and
8) they can use knowledge of native stress and tonal patterns.

The following sections address these points and discuss ways that the model could be fine-tuned to improve its performance.

7.2 Fine-tuning the Model

The kernel model focused on one aspect of speech variability that affects a listener's perception of intonation: variation in the pitch of the speaker's voice. In reality, other

aspects of speech may vary, such as speech rate. Speakers may produce the same sentence at different speeds, varying the rate within and between words. Consequently, corresponding syllables between two utterances of the same sentence would become misaligned. This misalignment could affect the relative timing of the sentence-type intonation cue between these two utterances. Among all three languages, the model is most sensitive to the timing of the statement and question cues in English because they are aligned with a stressed syllable. The model is less sensitive to the timing of the question cues in the tone languages because the Cantonese question rise occurs consistently at the end of the utterance and the Mandarin raised pitch range affects the entire utterance. Therefore, in addressing the timing issue in this section, I refer to English question cues only.

Figure 7.1 displays the F0 contours of "Ann is a teacher?" produced by a female (top) and a male (bottom) speaker. Compared with the male-produced contour, the female-produced contour is longer in duration and the timing of its nuclear accent (and question rise) is later in the utterance.[23] In addition to speaking rate, the relative position of the nuclear accent in the utterance also affects the timing of the English statement and question cues.

[23] Due to the aperiodicity of [tʃ] between 'tea…' [t^hi] and '…cher' [tʃɻ], there is a slight drop in F0 between the end of the voiced portion of [t^hi] and the start of the voiced portion of [tʃɻ]. This unvoiced gap makes it difficult to determine the final (steep) rise in the female utterance. It becomes more apparent in the interpolated contour shown in Figure 7.3(a) below. Although the production target of the final rise for this speaker might have been at the nuclear tone (L*), the F0 contour shows that the final rise starts in the final syllable.

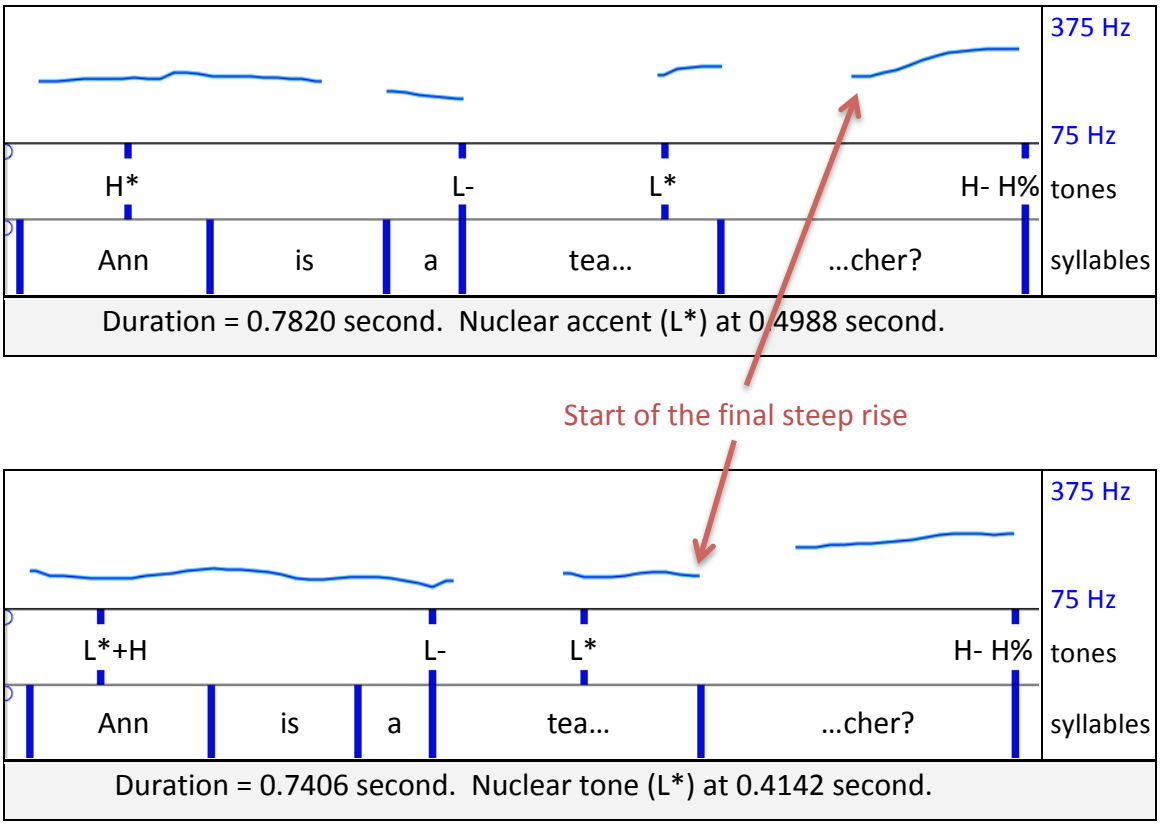


Figure 7.1. Variation in the timing of the nuclear accent in two different productions of the same question.

In Figure 7.1, the nuclear accent falls on the penultimate syllable of the utterance because the first syllable of "teacher" is stressed, whereas in Figure 7.2, the nuclear accent falls on the final syllable of the utterance because the final word "time" is stressed.

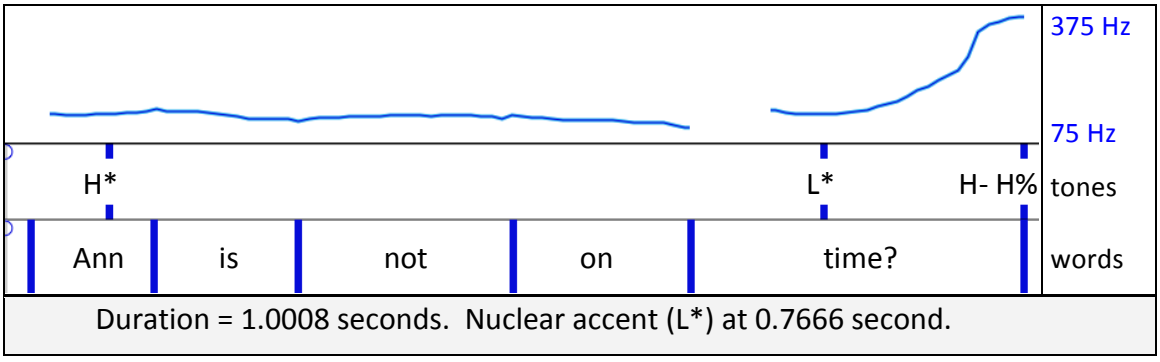


Figure 7.2. English question rise at a final stressed syllable.

The model's similarity calculation method handles time-scale variation between whole utterances, but not within or between words in the utterance. It uses F0 measurements extracted at relative, static time points of the utterance (i.e., at every 10% of the utterance). For example, Figure 7.3(a) displays the interpolated F0 contours of the two productions of "Ann is a teacher?" from Figure 7.1. Then Figure 7.3(b) displays these contours after they have been time-normalized at eleven equidistant time points. In both figures, the question rises between these two contours are misaligned. In the model, the misalignment of the sentence-type intonation cue between a token and an exemplar decreases the similarity between the two, and would therefore negatively affect the model's classification performance.

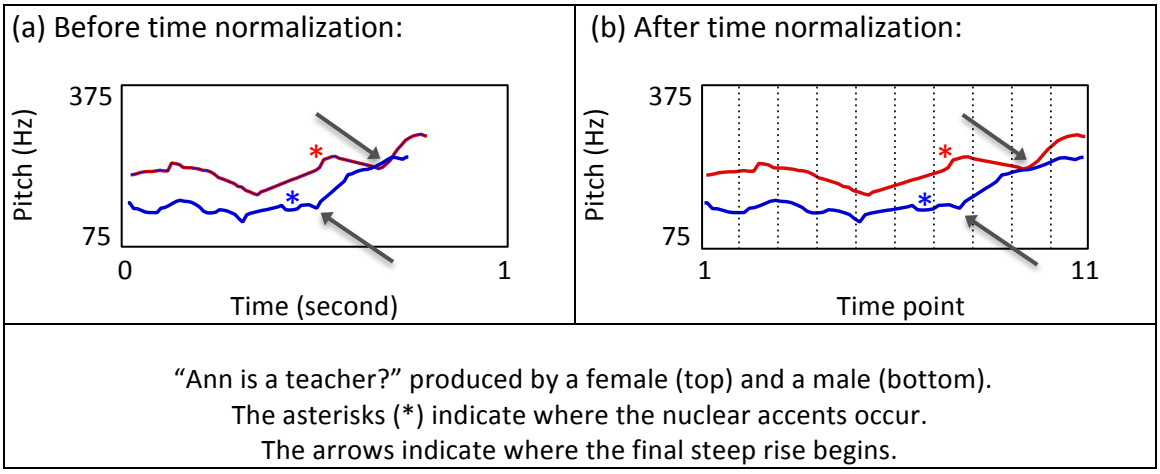


Figure 7.3. Misalignment of the question rise between a token (bottom) and an exemplar (top) in static time comparison.

One way to rectify this problem is to align the two utterances at specific points in the contours. At first thought, syllable boundaries appear to be suitable candidates for such 'time landmarks' in each utterance. However, sentences can have different syllable lengths. Even if two sentences have the same number of syllables, they are likely composed of different words. This means that the prominent stressed syllable (or the

nuclear syllable) probably differs between sentences, and consequently, so does the final rise or fall of the utterance. Since English signals statements with a final F0 fall and questions with a final F0 rise, a logical time-alignment point for English would be at the final maximum (for statements) or minimum (for questions) of the intonation contour.

To locate the final fall or rise in an utterance, my proposed method[24] would be to enable the model to compare a token with an exemplar iteratively, each time reducing the length of the comparison window for both utterances by a fixed amount (e.g., by one time point). Figure 7.4 illustrates this process using the two F0 contours from Figure 7.3(b). To keep the illustration simple, each utterance is compared using three window lengths only: the entire F0 contour, two-thirds of the F0 contour, and one-third of the F0 contour. The compared portions of the contours are displayed in the white area. For this illustration, I also assume that the F0 contour on the top is an exemplar 'in memory' and the F0 contour on the bottom is an incoming token.

[24] A well-known method for time-alignment of speech tokens that have similar patterns is dynamic time warping (DTW) (e.g., Müller, 2007; Rilliard, Allauzen, & Mareüil, 2011, for prosodic similarities; Sakoe & Chiba, 1978), which was introduced for application in automatic speech recognition. DTW matches a token with a target by stretching or compressing the token, while optimally minimizing the cost associated with the stretch and compression. In general, DTW is processing intensive and can potentially find more than one optimal solution. This thesis proposes a different alignment method.

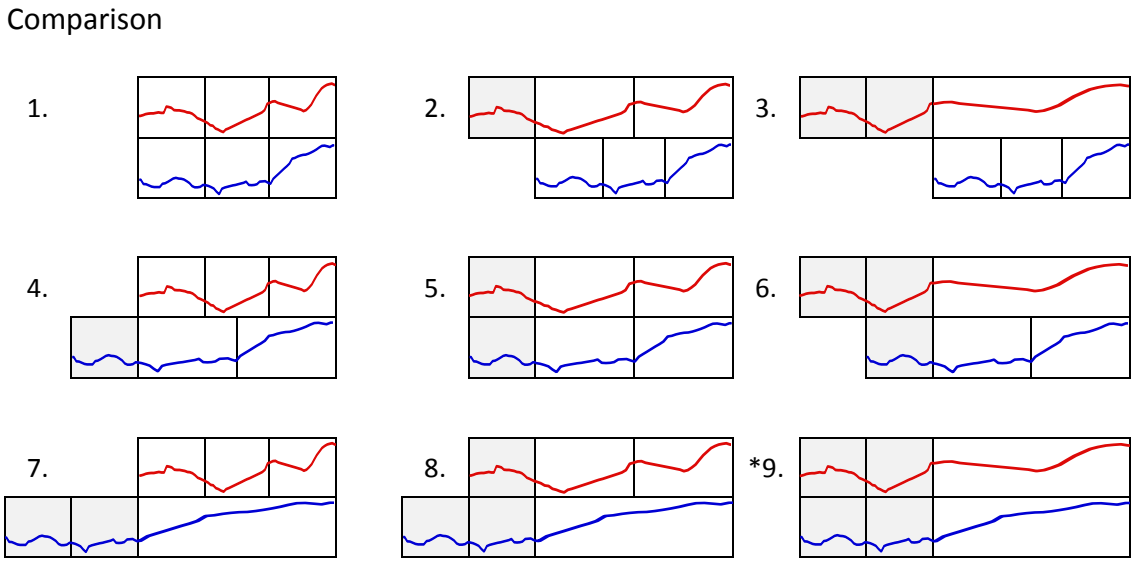


Figure 7.4. Dynamic time alignment process of an exemplar (top, red) with a token (bottom, blue), using three window lengths.

Initially, the entire contours of the token and the exemplar are compared (comparison 1). Then, as the process iterates through the exemplar (on the top), the starting point of the comparison window of the exemplar is shifted right by one-third of the length of the exemplar each time (comparisons 2 and 3). (As the compared portion of one utterance becomes smaller than that of the other, the smaller portion becomes stretched (or 'warped') relative to the other.) Similarly, as the process iterates through the token (on the bottom), the starting point of the comparison window of the token is shifted right by one-third of the length of the token each time (comparisons 4 and 7). Combining both iterative processes for the two utterances in this fashion would result in nine different comparisons. In each comparison, the model uses the same, general auditory distance calculation in (3.4) to determine how similar the compared portions of the utterances are to each other. At the end of the comparisons, the comparison with the smallest distance value would be considered to have the best aligned contours.

In this example, comparison 9 has the best-aligned contours. Due to the relatively large window step size (one-third of the length of the F0 contour), the nuclear accent of the exemplar happens to align with the final rise of the token in this case. If the window step size was smaller (e.g., one-fifth of the F0 contour), then the likelihood of a better alignment is greater, as shown in Figure 7.5.

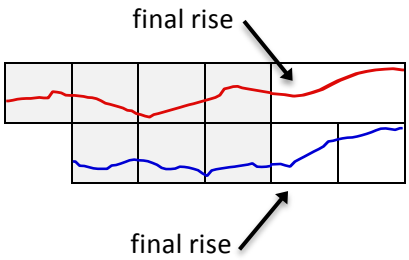


Figure 7.5. Dynamic time alignment of an exemplar (top, red) with a token (bottom, blue) using five window lengths.

The proposed time-alignment method would align only the final fall or rise and not the entire F0 contour. As the results of the human performance and the model's performance in the classification task indicate, both the computer model and the human listeners were significantly less sensitive to stimuli that excluded the final syllable than to stimuli that included it (Section 6.4.2). These results suggest that F0 is a less salient cue prior to the penultimate syllable (or the final F0 rise), so—without assuming whether or not human listeners actually align the F0 contour prior to the final fall/rise between utterances—I propose not to align the less salient portion of the utterance (until perhaps after testing the model with the alignment of the sentence-type cues).

As discussed in Section 6.4.4, the English listeners were able to identify the sentence type even if only part of the question cue was available from the stimulus. This suggests that the human listeners were comparing the utterance fragments with the whole utterances, which contained the entire final statement/question cue. The kernel model was

comparing the tokens with the exemplars 'in memory' that were of the same stimulus type as the tokens, but it needs to compare all tokens with the whole exemplars 'in memory' as well. However, since part of the utterance is missing from the fragment, it cannot be normalized by time with the whole utterance. The model can use a variant of the time alignment method above to simulate this type of comparison.

Figure 7.6 illustrates this alignment process using the F0 contour of the utterance "Ann is not on time?" in Figure 7.2 as one exemplar (displayed at the top of the figure) and its corresponding statement utterance "Ann is not on time." as another exemplar (displayed at the bottom of the figure). Two of the fragments from the question utterance serve as two new tokens: "Ann is not on" (NoLast) and "time" (Last). There are four processes shown in Figure 7.2: 1) the NoLast stimulus (on the left) compared with the question exemplar, 2) the NoLast stimulus compared with the statement exemplar, 3) the Last stimulus (on the right) compared with the question exemplar, and 4) the Last stimulus compared with the statement exemplar.

For this illustration, seven time points are used. In each process, the model could first align the start and end of the fragment with time points 1 and 2 of the whole utterance, respectively. Then the model could compare the fragment with the whole utterance successively by shifting the end of the fragment forward by one time point of the whole utterance each time until it reaches the last time point of the whole utterance (in comparison 6). The model could then continue to compare the fragment with the whole utterance successively by shifting the start of the fragment forward by one time point of the whole utterance at each comparison step. In this example, the best alignment for the NoLast stimulus would be comparison 4 with the question exemplar, and the best

alignment for the Last stimulus would be comparison 11 with the question exemplar. The best match for the NoLast and Last stimuli would be the question exemplar, and not the statement exemplar.

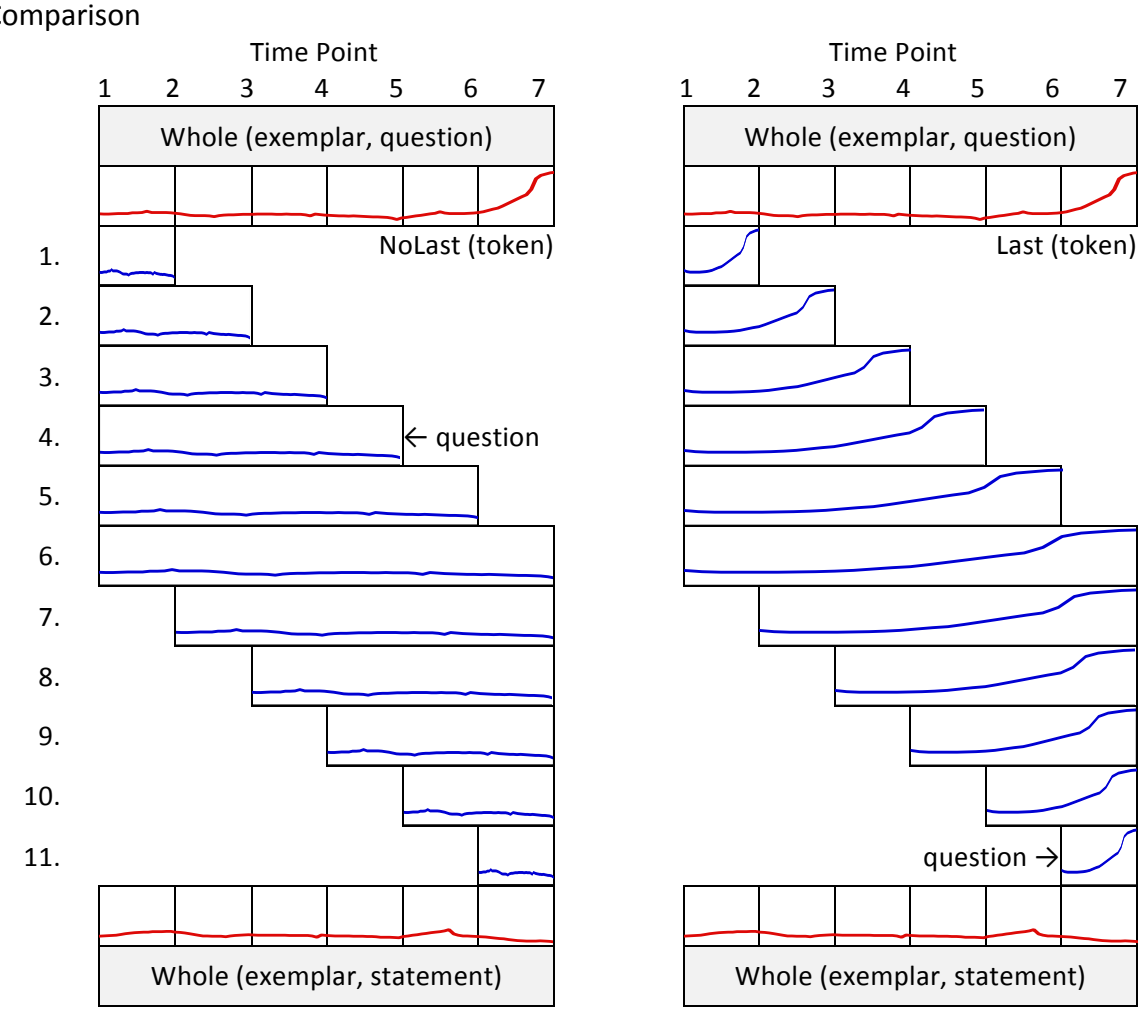


Figure 7.6. Alignment of two fragments with a whole utterance (question, top; or statement, bottom) through 11 comparisons.

As Table 5.4 shows, the Cantonese and Mandarin listeners performed significantly better on whole utterances than on the final syllable alone. This suggests that there are still some identifiable cues in the early part of the utterance. This is especially true for Mandarin, where questions are signaled by a higher pitch than that of a statement throughout the utterance.

In Mandarin, the lack of a significant difference in the listeners' performance on NoLast and Last (Table 5.4) suggests that the final F0 rise or fall is a moderately salient cue in the language. Since the elevated pitch in Mandarin questions co-occurs with pitch range expansion—as the F0 ranges of the speakers' productions of the stimulus type First in Figure 7.7 show—pitch range measurements could be used in the similarity calculation to accommodate Mandarin's question intonation.

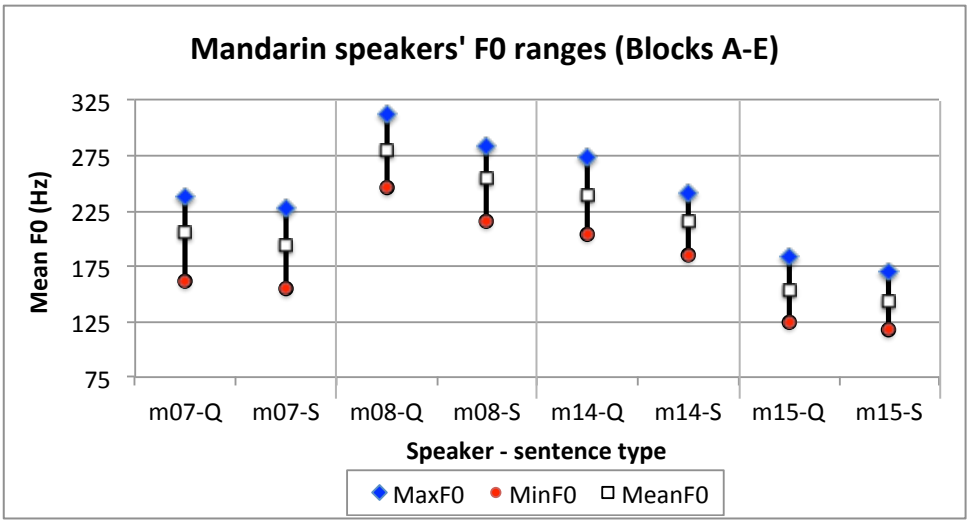


Figure 7.7. F0 ranges of the Mandarin speakers' production of stimulus type First, averaged over all five blocks.

The raised pitch range of Mandarin questions is a global cue, rather than a local cue. For example, Figure 7.8 shows the statement and question pair from Figure 2.9. The red dotted lines show the gradual fall in F0 in the utterance at the top and the gradual rise in F0 (a salient global cue for questions) in the utterance at the bottom.

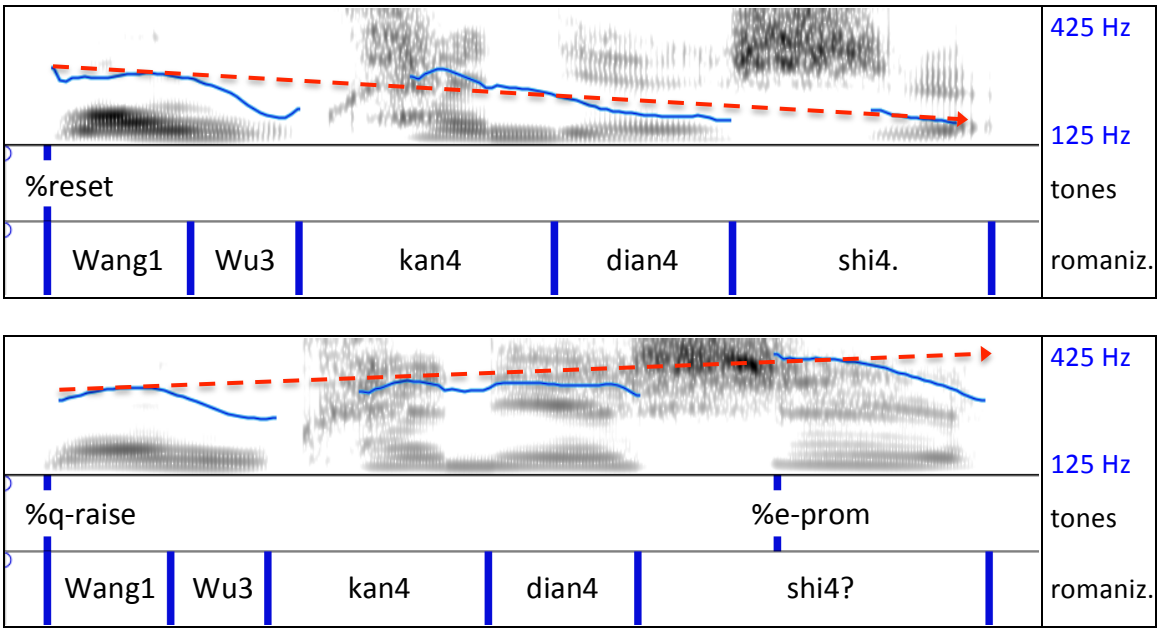


Figure 7.8. Intonation cues (red dotted lines) for a Mandarin statement and question: 'Wang Wu watches TV'.

Using F0 values at individual time points would likely capture more of the local, tonal variation. Instead of working only with the F0 values at eleven equidistant time points of the utterance, the auditory distance between a token and an exemplar could be calculated using the mean F0 (meanF0) values of the intonation contours between every two successive time points. That is, meanF0_1 is the mean F0 of the intonation contour between time points 1 and 2, meanF0_2 is the mean F0 of the intonation contour between time points 2 and 3, meanF0_3 is the mean F0 of the intonation contour between time points 3 and 4, and so on. The formula for the auditory distance between token *i* and exemplar *j* is shown in (7.1).

(7.1)

$$d_{ij} = \left[\sum_{n=1}^{p-1}\left(\text{meanF0}_{ni} - \text{meanF0}_{nj}\right)^2\right]^{1/2}, \text{ where } p = 11$$

Fine-tuning the model may improve its performance but may also complicate it with language-specific criteria, ultimately requiring it to separate into independent language models. To retain a single-model design, the preference would be to combine both the mean F0 and F0 height comparisons in the same model for all three languages, rather than applying mean F0 only in the Mandarin model and F0 height for the English and Cantonese models. This integrated design would have the potential for modeling second language intonation perception as well. For example, in fieldwork eliciting non-native Cantonese speech from a native Mandarin speaker (Chow, 2016), this language consultant produced Cantonese echo questions with both the Cantonese and Mandarin echo question cues in the intonation, that is, with a high final F0 rise characteristic of Cantonese and an elevated pitch characteristic of Mandarin. In principle, combining the similarity calculation formulas in (3.4) and (7.1), as shown in (7.2), would work with these non-native Cantonese utterances. This hypothesis, of course, needs to be tested, but second language intonation perception is beyond the scope of this thesis.

(7.2)

$$d_{ij} = \left[\sum_{n=1}^{p}\left(\mathrm{F0}_{ni} - \mathrm{F0}_{nj}\right)^2 + \sum_{n=1}^{p-1}\left(\mathrm{meanF0}_{ni} - \mathrm{meanF0}_{nj}\right)^2\right]^{1/2}, \text{ where } p = 11$$

7.3 Additional Mechanisms for the Model

In perceiving native intonation in statements and echo questions, listeners of a language tend to pay more attention to the auditory properties that provide the sentence-type intonation cues in their language. To compensate for this language specificity, my model could apply attention weights (Johnson, 1997; Nosofsky, 1988) to the auditory properties

in its similarity calculation. Adding an attention weight w_m to each auditory property m in the calculation of auditory similarity d_{ij} in (7.2) generates the formula in (7.3).

(7.3)

$$d_{ij} = \left[\sum_{n=1}^{p} w_n \left(\mathrm{F0}_{ni} - \mathrm{F0}_{nj} \right)^2 + \sum_{n=1}^{p-1} w_{p+n} \left(\mathrm{meanF0}_{ni} - \mathrm{meanF0}_{nj} \right)^2 \right]^{1/2}, \qquad \text{where } p = 11$$

Johnson (1997) applied attention weights differentially across the auditory properties in his exemplar-based model of vowel perception by using a simulated annealing algorithm (Kirkpatrick, Gelatt, & Vecchi, 1983) to optimize its performance. The attention weights that were determined by the simulated annealing algorithm depended on the type of simulation. For example, for gender identification, F0 would receive a greater attention weight than F1 and F2.

Although a simulated annealing algorithm could help my computer model to perform better, it could also obscure the reasons why the model was not performing as well as the human listeners. The performance differences between the computer model and the human listeners could provide direction on how to improve the model's performance or to make it a better model of human behaviour, in general. For example, the perceptual results in Chapters 5 and 6 suggest that, for English and Cantonese, the model would need more weight on F0 height towards the end of the utterance than in the earlier part of the utterance. For Mandarin, since the elevated pitch is gradual from the start to the end of the utterance (Figure 5.2), the relative weights of the mean F0s should increase from the first time interval to the last interval as well. Comparisons of the F0 values may not be necessary, so their weights could be set to zero. In general, the

estimates of the SpC could be used by the model to determine the relative weights of the (mean) F0 values for all three languages. The higher the estimated probability (suggesting likelihood of better performance), the greater the weight of the (mean) F0 value.

Finally, extra dimensionality could be added to the similarity calculation by factoring in additional cues. For example, final lengthening can be a secondary question cue, so the duration of the final syllable, dur, can be added to the formula in (7.3), as shown in (7.4).

(7.4)

$$d_{ij} = \left[\sum_{n=1}^{p} w_n\left(\mathrm{F0}_{ni} - \mathrm{F0}_{nj}\right)^2 + \sum_{n=1}^{p-1} w_{p+n}\left(\mathrm{meanF0}_{ni} - \mathrm{meanF0}_{nj}\right)^2 + w_{2p}\left(\mathrm{dur}_i - \mathrm{dur}_j\right)^2\right]^{1/2}, \quad \text{where } p = 11$$

7.4 Considerations for a Generalized Model

In fine-tuning and enhancing the kernel model to better simulate human listeners' performance in identifying statements and questions based on intonation, the aim for a generalized intonation perception model remains. A generalized model that can deal with multiple languages can reveal cross-linguistic differences in how native listeners process different sentence intonation patterns in different languages. However, since intonation can interact with other elements of prosody, such as lexical tones, at some point in the development, this model would need to deal with language-specific features. As the comparative results between the computer model and the human listeners suggest, experience with native stress or tonal patterns contributes to the better performance of the

human listeners. For example, in Mandarin, the computer model did not perform significantly differently from the human listeners on final syllables that carried a falling tone, but did perform significantly worse than the human listeners on final syllables that carried a low tone. As shown in Figure 5.4, the Mandarin low tone has two allotones: the allotone [21] is realized at the end of statements and the allotone [214] is realized at the end of questions. In this case, the model would need to be trained on Mandarin tones that appear in final syllables of statements and questions as well, in order to be able to deal with tonal variation. (One way to simulate this training could be to store isolated utterance-final tones for the four tonal categories in Mandarin as exemplars in the model's 'memory', as shown in the Tone 3 exemplar cloud in Figure 7.9. In reality, it is likely that a native Mandarin-speaking child would have experienced statements and questions expressed by a single Mandarin tone, e.g., *ma3*? 'horse'.) In addition, in categorizing a token, the model would need to categorize both its sentence type and its final tone. Figure 7.9 displays how the model could simultaneously categorize tokens by sentence type and final tone such that both the intonation and lexical tone information would be available for categorizing subsequent tokens. This method of implementation requires training and testing of additional categories, but does not limit the model with language specificities. In principle, it can handle sociolinguistic variation, such as uptalk, using additional categories in the same manner as for the Mandarin tones.

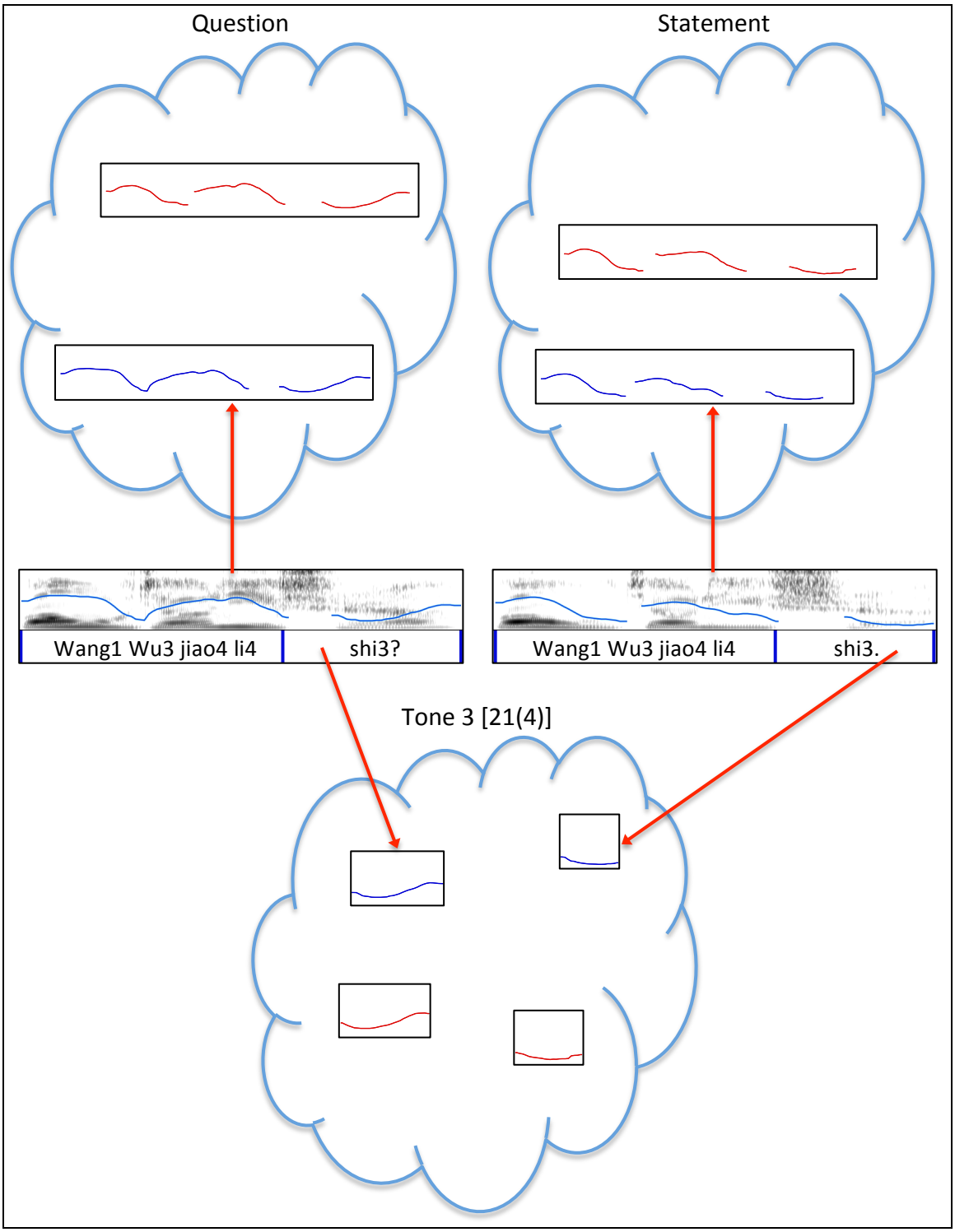


Figure 7.9. Categorization of two tokens of "Wang Wu teaches history" (middle) by sentence-type intonation (top) and final tone (bottom).

The kernel model has demonstrated that it can successfully categorize statements and questions, based on their sentence-type intonation. In everyday communication, there are other sentence types with different intonation patterns, such as imperative and exclamatory sentences (Wells, 2006), which this exemplar-based model has yet to address. In principle, this model could categorize these sentence types using the same similarity calculation and auditory properties as it did for the statements and questions. With increased variation in sentence types, however, it is expected that the performance of the computer model would decline. Increased experience with naturally produced sentences may help, but if the sentence-type intonation cues are very similar, additional auditory properties may be needed in order to enable the model to reach human levels of performance in intonation categorization. Finally, in order for the model to accommodate additional languages, it may need to address other elements of intonation, such as pitch range or pitch accents.

7.5 Summary

This chapter has discussed ways to improve the performance of the kernel model by adjusting and increasing its functions to better simulate the human process of categorizing statements and questions. These suggestions were inspired by the differences between the computer model's performance and the human listeners' performance in the classification task. While aiming towards a generalized model, I propose that the model use dynamic time alignment to handle timing variation, mean F0 auditory properties to capture global question cues, attention weights to focus on the portion of the utterance that contains the salient sentence-type cue, and additional

categories to simulate tonal and other language-specific knowledge that is relevant to the sentence-type identification task. The concluding chapter that follows summarizes the general findings of this thesis and offers some future directions for the modeling of intonation perception.

Chapter 8: Conclusion

8.1 Findings

This thesis has proposed an exemplar-based model to account for native listeners' categorization of sentence intonation in English, Cantonese, and Mandarin. The exemplar theory of speech perception (Goldinger, 1998; Johnson, 1997) addresses the question of how human listeners can cope with the massive, inherent variability in speech. According to exemplar theory, listeners retain the fine acoustic details of their speech experience in stored exemplars in memory. During speech processing, this detailed information enables listeners to categorize a new instance of an utterance, based on its overall similarity with the exemplars for that category in memory.

Both an exemplar-based computational model and human listeners classified statements and questions produced by native speakers in English, Cantonese, and Mandarin. They were presented with these utterances, gated in five forms: Whole, NoLast, Last, Last2, and First. The computer model simulated an exemplar-based process of categorization that was based solely on a comparison of the F0 values between a token and the exemplars 'in memory'. The computer model correctly classified the statements and questions in all three languages at better than chance rates, suggesting that F0 is a salient cue for identifying the sentence types in these languages. Similar to Johnson's (1997) study on the categorization of vowels, this exemplar-based model categorized sentence intonation without normalizing F0 for each speaker's voice. The result of its performance demonstrated that an exemplar-based model is a promising tool for investigating how human listeners process variation in intonation in speech.

Compared to the human listeners, the computer model performed significantly worse in the classification task, suggesting that the human listeners might be using other cues, besides F0, for identifying statements and questions in these languages. This result also indicates that human listeners' experience with their native language intonation system (i.e., exemplars of statement and question intonation in memory) helped to improve their performance.

In Cantonese, both the model and the listeners performed similarly on the final two syllables. However, the model performed significantly worse on whole utterances than on the final two syllables, while the listeners performed similarly on both stimulus types. This result suggests that the listeners might pay less attention to parts of the utterance that do not contain the salient cue (i.e., prior to the penultimate syllable). Additionally, these listeners responded to the final syllables with more bias towards questions than statements, compared to the computer model. This response behaviour suggests that the listeners' rich experience with the question cue in their native language might have reduced their focus of attention on the acoustic cues of these stimuli.

In English, the model performed significantly better when the stimuli contained the entire question rise, rather than just part of the rise, whereas the listeners did not perform significantly differently on stimuli that contained either the entire question rise or just part of it. The fact that the listeners were able to recover part of the missing cue suggests that they might be storing whole intonation patterns in memory and that they were able to match potential intonation patterns to the relevant parts of whole intonation contours stored in memory.

In Mandarin, the model did not perform significantly differently from the human listeners on final syllables that carried a high or falling tone, but it performed significantly worse than these listeners on final syllables that carried a rising or low tone. This evidence suggests that the listeners were using their experience of native tones to help them categorize the sentence types. The fact that these listeners performed better on sentences that end in one tone than on sentences that end in another tone suggests that they could not have used a single, abstract representation of the question (or statement) intonation pattern to categorize all of these sentences. In summary, the comparative results between the performance of the computer model and the human listeners suggest that human listeners store whole intonation patterns in memory and use the detailed information from these patterns, selectively, to categorize the sentence type of new utterances.

The model was significantly more sensitive to the statement/question distinction in English and Cantonese than in Mandarin, similar to the human listeners. This evidence suggests that the question cue (a high F0 rise) in English and Cantonese might be easier for both the computer model and the native listeners to detect than the question cue (a raised pitch range) in Mandarin. However, the computer model was considerably less sensitive to Mandarin than the Mandarin listeners were, suggesting that the model, in its current form, does not detect Mandarin's global question cue well. Nevertheless, the model demonstrated that it could, in general, account for the differences in the statement and question intonation patterns of all three languages.

8.2 Contribution

As far as I know, this study is the first that extended exemplar theory to account for the human perception of intonation in statements and questions in English, Cantonese, and Mandarin. The proposed model has demonstrated that it is feasible to use computer modeling as a scientific means to explore the process of categorizing statement and question intonation patterns by human listeners. The findings of this study contribute to ongoing research on speech perception and provide insight into how human listeners process variation in intonation across languages. The acoustic analysis of the different intonation patterns across all three languages and the perceptual responses of the native listeners in the sentence-type identification task advance the knowledge of cross-linguistic similarities and differences in signaling questions and statements with intonation cues, as well as the understanding of the interaction between lexical tones and sentence intonation in Cantonese and Mandarin.

8.3 Limitations

As a pioneering study on the application of exemplar theory to the analysis of the perception of intonation in English, Cantonese, and Mandarin, this study had some necessary limitations. First of all, both the computer model and the human listeners were presented with utterances that were gated at sentence or word boundaries in order to identify how much intonational information listeners could get from each gated portion of an utterance. In normal conversations, the sentences that listeners hear are usually continuous, either produced by a single speaker or multiple speakers. However, previous studies (e.g., Jusczyk, Houston, & Newsome, 1999; Mattys, Jusczyk, Luce, & Morgan,

1999; Yip, 2017) have found that listeners (e.g., infants) can use their knowledge of native phonological patterns to segment words. Secondly, the eleven F0 values that the model used to compare the similarity between a token and an exemplar were time normalized at eleven equidistant time points. This implementation simplified the processing of the computer model, but does not imply that the human listeners take all of the same steps in the process of categorizing similar tokens. Lastly, interpolation was necessary to fill the unvoiced gaps in the intonation contours to enable the model to extract F0 values at equidistant time points of the contour. It is unknown how human listeners process unvoiced gaps in an intonation contour.

8.4 Future Directions

Given the promising results of this initial, cross-linguistic study on the perception of intonation in statements and questions by human listeners and an exemplar-based model, the next step would be to repeat the simulations of the kernel model using sentences produced by all of the speakers in the production study to investigate the exemplar effects of increased experience (i.e., more exemplars) and variability (i.e., more speakers) on the model's performance in categorizing these utterances. In particular, would increased experience improve performance on one language more than another?

Another possible future direction would be to test human listeners who are not native listeners of the language under study. Native listeners have prior experience with their native language's intonation system, which the model lacks, so it would be logical to conduct the same perception study on naïve human listeners, who have no prior knowledge or experience with the target language. The difference in performance

between the naïve and native listeners would help to determine how much performance is affected by experience with the target language; the potential difference in performance between the naïve, human listeners and the computational model would help to determine how much performance is affected by the native language experience of the naïve listeners.

Furthermore, the proposed model is intended to become a generalized model that can account for different intonation systems. A good test of the model's generalizability would be to repeat this study with another language whose intonation system differs from English, Cantonese, and Mandarin. A recommended language would be an African language that expresses questions with a falling intonation, known as lax prosody (Rialland, 2009). This type of question intonation seems to be dependent on F0 height, but its directionality is the opposite of that in the question intonation of English and Cantonese. Additionally, there is a general tendency for statements to fall (Cohen, Collier, & 't Hart, 1982; Pike, 1945; Vaissière, 1983), which suggests that the F0 difference between statements and questions could be potentially less salient for languages that signal questions with a falling intonation, rather than a rising intonation.

Finally, Chapter 6 has provided explanations for some of the reasons why the model was not performing as well as the human listeners, and Chapter 7 has proposed adjustments and enhancements to the model to address these issues. Further work could include simulations with the enhanced version of the model to determine if those enhancements actually improve the model's performance. If so, the enhanced model could be used to further address the issue of how human listeners process variability in intonation perception, in an exemplar-theoretic framework.

References

Adams, C., & Munro, R. R. (1978). In search of the acoustic correlates of stress: fundamental frequency, amplitude, and duration in the connected utterance of some native and non-native speakers of English. *Phonetica, 35*(3), 125-156.

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology. Trends in Linguistics Studies and Monographs 102*. Berlin: Mouton de Gruyter.

Beckman, M. E. (1986). *Stress and non-stress accent*. Dordrecht: Foris.

Beckman, M. E., & Hirschberg, J. (1999). The ToBI annotation conventions. Retrieved from http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S. A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 9-54). New York: Oxford University Press.

Boersma, P., & Weenink, D. (2013, May 30). *Praat: doing phonetics by computer*. [Computer application, version 5.3.51]. Retrieved from http://www.praat.org

Bolinger, D. L. (1979). Intonation across languages. In J. Greenberg (Ed.), *Universals of language: vol. 2. Phonology* (pp. 471-524). Stanford: Stanford University Press.

Brown, G., Anderson A., Shillcock, R., & Yule, G. (1984). *Teaching talk.* Cambridge: Cambridge University Press.

Bruce, G. (1982). Textual aspects of prosody in Swedish. *Phonetica, 39*, 274–287.

Calhoun, S., & Schweitzer, A. (2012). Can intonation contours be lexicalised? Implications for discourse meanings. In G. Elordieta & P. Prieto (Eds.), *Prosody*

and Meaning (Trends in Linguistics): vol. 25 (pp. 271-327). Munchen: Walter de Gruyter.

Chao, Y.-R. (1947). *Cantonese primer*. Cambridge: Harvard University Press.

Chao, Y.-R. (1948). *Mandarin primer: An intensive course in spoken Chinese*. Cambridge: Harvard University Press.

Chow, U. Y. (2016). L2 transfer of stress, tones, and intonation from Mandarin: A case study. *Calgary Working Papers in Linguistics, 29*, 19-40.

Chow, U. Y., & Winters, S. J. (2015). Exemplar-based classification of statements and questions in Cantonese. *Proceedings of the 18th International Congress of Phonetic Sciences*. Paper number 0987.1-5.

Chow, U. Y., & Winters, S. J. (2016). The role of the final tone in signaling statements and questions in Mandarin. *Proceedings of the 5th International Symposium on Tonal Aspects of Languages*, 167-171.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(3), 521-533.

Ciocca, V., & Lui, J. (2003). The development of lexical tone perception in Cantonese. *Journal of Multilingual Communication Disorders, 1*, 141-147.

Cohen, A., Collier, R., & ‘t Hart, J. (1982). Declination: construct or intrinsic feature of speech pitch? *Phonetica, 39*, 254-273.

Di Gioacchino, M., & Jessop, L. C. (2011). Uptalk-Towards a quantitative analysis. *Toronto Working Papers in Linguistics*, *33*(1).

Duanmu, S. (2007). *Phonology of Standard Chinese*. Oxford: Oxford University Press.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech weave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360-385). Hillsdale: Erlbaum.

Fant, G. (1970). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations: vol. 2*. The Hague: De Gruyter Mouton.

Fant, G. (1972). Vocal tract wall effects, losses, and resonance bandwidths. *Speech Transmission Laboratory Quarterly Progress and Status Report, 2*(3), 28-52.

Flynn, C.-Y.-C. (2003). *Intonation in Cantonese*. *LINCOM Studies in Asian Linguistics 49*. Muenchen: LINCOM GmbH.

Fok-Chan, Y. Y. (1974). *A perceptual study of tones in Cantonese*. Hong Kong: University of Hong Kong Press.

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, *1*(2), 126-152.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166-1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. *Proceedings of the 16th International Congress of Phonetic Sciences*, 49-54.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics, 31*(3), 305-320.

Goldsmith, J. (1976). *Autosegmental phonology*. PhD dissertation, MIT.

Goldsmith, J. (1990). *Autosegmental and metrical phonology*. Cambridge: Basil Blackwell.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics, 28*, 267-283.

Gu, W., Hirose, K., & Fujisaki, H. (2005). Analysis of the effects of word emphasis and echo questions on F0 contours of Cantonese utterances. *Proceedings of the 9th European Conference on Speech Communication and Technology*, 1825-1828.

Gussenhoven, C., & Chen, A. (2000). Universal and language-specific effects in the perception of question intonation. *Proceedings of the 6th International Conference on Spoken Language Processing*, 91-94.

Hartman, L. M. (1944). The segmental phonemes of Peiping dialect. *Language, 20*, 28-42.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528.

Hirst, D. (1983). Structures and categories in prosodic representations. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 93-109). Berlin: Springer.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145-165). San Diego: Academic Press.

Johnson, K. (2005). Speaker normalization in speech perception. In D. Pisoni & R. Remez (Eds.), *The handbook of speech perception* (pp. 363-389). Malden: Blackwell.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34*(4), 485-499.

Jun, S. A. (Ed.). (2005). *Prosodic typology: The phonology of intonation and phrasing*. New York: Oxford University Press.

Jun, S. A. (Ed.). (2014). *Prosodic typology II: The phonology of intonation and phrasing: vol. 2*. Oxford: Oxford University Press.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*(3), 159-207.

Kirchner, R., Moore, R. K., & Chen, T. Y. (2010). Computing phonological generalization over real speech exemplars. *Journal of Phonetics, 38*(4), 540-547.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671-680.

Ladd, D. R. (2008). *Intonational phonology (2nd ed.)*. Cambridge: Cambridge University Press.

Ladefoged, P. (1982). *A course in phonetics*. San Diego: Harcourt Brace Jovanovich Publishers.

Lehiste, I., & Meltzer, D. (1973). Vowel and speaker identification in natural and synthetic speech. *Language and Speech, 16*, 356-364.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (Eds.). (2016). *Ethnologue: Languages of the World (19th ed.)*. Dallas: SIL International. Retrieved from http://www.ethnologue.com

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431– 461.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus variables in perception of unvoiced stop consonants. *American Journal of Psychology, 65,* 497-516.

Liu, F., Surendran, D., & Xu, Y. (2006). Classification of statement and question intonations in Mandarin. *Proceedings of the 3rd International Conference on Speech Prosody*. Paper 232.

Ma, J. K.-Y., Ciocca, V., & Whitehill, T. L. (2006). Effect of intonation on Cantonese lexical tones. *Journal of the Acoustical Society of America, 120*(6), 3978-3987.

Ma, J. K.-Y., Ciocca, V., & Whitehill, T. L. (2011). The perception of intonation questions and statements in Cantonese. *Journal of Acoustical Society of America, 129(2)*, 1012–1023.

Macmillan, N.A., Creelman, C.D. (2005). *Detection theory: A user's guide (2nd ed.)*. New York: Cambridge University Press.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*(2), 391-409.

Masters, T. (1995). *Advanced algorithms for neural networks: A C++ sourcebook*. New York: Wiley.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*(4), 465-494.

Müller, M. (2007). Dynamic time warping. In *Information retrieval for music and motion* (pp. 69-82). Berlin: Springer.

Mok, P. P. K., Zuo, D., & Wong, P. W. Y. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change, 25*, 341-370.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39-57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 700–708.

Nygaard, L.C. (2005). The integration of linguistic and non-linguistic properties of speech. In D. Pisoni & R. Remez (Eds.), *Handbook of speech perception* (pp. 390–414). Malden, MA: Blackwell.

Paeschke, A. (2004). Global trend of fundamental frequency in emotional speech. *Proceedings of the 2nd International Conference on Speech Prosody*, 671-674.

Peng, S.-H., Chan, M. K. M., Tseng, C.-Y., Huang, T., Lee, O. J., & Beckman, M. E. (2005). Towards a Pan-Mandarin system for prosodic transcription. In S.-A. Jun

(Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 230-270). New York: Oxford University Press.

Peterson, G., & Barney, H. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*(2), 175-184.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. Ph.D. dissertation, M.I.T.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), Frequency effects and emergent grammar (pp. 137-157). Amsterdam: John Benjamins.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonation contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. Pollack (Eds.), *Plans and intentions in communication and discourse* (pp. 271-311). Cambridge: MIT Press.

Pike, K. L. (1945). *The intonation of American English*. Ann Arbor: University of Michigan Press.

Raphael, L. J., Borden, G. J., & Harris, K. S. (2011). *Speech science primer: Physiology, acoustics, and perception of speech (6th ed.)*. Baltimore: Lippincott Williams & Wilkins.

Reetz, H., & Jongman, A. (2009). *Phonetics: Transcription, production, acoustics, and perception*. Oxford: Wiley-Blackwell.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In L. Liu & M. T. Zsu (Eds.), *Encyclopedia of database systems: vol. 6* (pp. 532-538). Berlin: Springer.

Rialland, A. (2009). The African lax question prosody: Its realisation and geographical distribution. *Lingua, 119*(6), 928-949.

Rilliard, A., Allauzen, A. & Boula de Mareüil, P. (2011). Using dynamic time warping to compute prosodic similarity measures. *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 2021-2024.

Ryalls, J., & Lieberman, P. (1982). Fundamental frequency and vowel perception. *Journal of the Acoustical Society of America, 72*, 1631-1634.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1)*, 43-49.

Shih, C. (2000). A declination model of Mandarin Chinese. In A. Botinis (Ed.), *Intonation: Analysis, modelling and technology: vol. 15* (pp. 243-268). Dordrecht: Springer.

Trimble, J. C. (2013). Perceiving intonational cues in a foreign language: Perception of sentence type in two dialects of Spanish. In C. Howe (Ed.), *Selected Proceedings of the 15th Hispanic Linguistics Symposium* (pp. 78-92). Somerville: Cascadilla Proceedings Project.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381-402). New York: Academic Press.

Urbanek, S., Bibiko, H.-J., & Iacus, S. M. (2013, September 10). *R for Mac OS X GUI*. The R Foundation for Statistical Computing. [Computer application, version 3.0.1]. Retrieved from http://www.R-project.org

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 53-66). Heidelberg: Springer-Verlag.

Van Heuven, J. V. (2004). Planning in speech melody: Production and perception of downstep in Dutch. In H. Quené & J. V. Van Heuven (Eds.), *On speech and language: Studies for Sieb G. Nooteboom* (pp. 83–93). The Netherlands: LOT Occasional series by Utrecht University.

Vance, T. J. (1976). An experimental investigation of tone and intonation in Cantonese. *Phonetica, 33*, 368-392.

Walsh, M., Schweitzer, K., & Schauffer, N. (2013). Exemplar-based pitch accent categorisation using the Generalized Context Model. *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, 258-262.

Warren, P. (2005). Patterns of late rising in New Zealand English: Intonational variation or intonational change? *Language Variation and Change, 17*, 209-230.

Wedel, A. (2004). Category competition drives contrast maintenance within an exemplar-based production/perception loop. *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology*, 1-10.

Wells, J. C. (2006). *English intonation: An introduction*. Cambridge: Cambridge University Press.

Wong, W. Y. P., Chan, M. K. M., & Beckman, M. E. (2005). An autosegmental-metrical analysis and prosodic annotation conventions for Cantonese. In S. A. Jun (Ed.),

Prosodic typology: The phonology of intonation and phrasing (pp. 271-300). New York: Oxford University Press.

Xu, B. R., & Mok, P. (2011). Final rising and global raising in Cantonese intonation. *Proceedings of the 17th International Congress of Phonetic Sciences*, 2173-2176.

Xu, Y., & Wang, Q. E. (1997). What can tone studies tell us about intonation? *Intonation: Theory, Models, and Applications*, 337-340.

Yip, M. C. (2017). Probabilistic phonotactics as a cue for recognizing spoken Cantonese words in speech. *Journal of Psycholinguistic Research, 46*(1), 201-210.

Yuan, J. (2011). Perception of intonation in Mandarin Chinese. *The Journal of the Acoustical Society of America*, *130*(6), 4063-4069.

Yuan, J., & Shih, C. (2004). Confusability of Chinese intonation. *Proceedings of the 2nd International Conference on Speech Prosody*, 131-134.

Yuan, J., Shih, C., & Kochanski, G. P. (2002). Comparison of declarative and interrogative intonation in Chinese. *Proceedings of the International Conference on Speech Prosody*, 711-714.

Appendix A: Stimuli

A.1 English Stimuli

Block A (target sentences = 5 syllables long)

1. A: Who is Ann?
   B: Ann is a teacher.
   A: Ann is a teacher?
   B: Yes, Ann is a teacher.

2. A: What does Ann teach?
   B: Ann teaches history.
   A: Ann teaches history?
   B: Yes, Ann teaches history.

3. A: Why isn't Ann here?
   B: Ann is not on time.
   A: Ann is not on time?
   B: Yes, Ann is not on time.

4. A: Does Ann like to watch films?
   B: Ann likes to watch films.
   A: Ann likes to watch films?
   B: Yes, Ann likes to watch films.

Block B (target sentences = 7 syllables long)

1. A: Who is Mary?
   B: Mary is a good dentist.
   A: Mary is a good dentist?
   B: Yes, Mary is a good dentist.

2. A: What is Mary buying?
   B: Mary is buying a chair.
   A: Mary is buying a chair?
   B: Yes, Mary is buying a chair.

3. A: Why would Mary know Neil?
   B: Mary is Neil's lovely aunt.
   A: Mary is Neil's lovely aunt?
   B: Yes, Mary is Neil's lovely aunt.

4. A: Has Mary forgotten Al?
   B: Mary has forgotten Al.
   A: Mary has forgotten Al?
   B: Yes, Mary has forgotten Al.

Block C (target sentences = 9 syllables long)

1. A: Who is Alice?
   B: Alice is an old high school friend's Mom.
   A: Alice is an old high school friend's Mom?
   B: Yes, Alice is an old high school friend's Mom.

2. A: What did Alice do?
   B: Alice went horse riding with a friend.
   A: Alice went horse riding with a friend?
   B: Yes, Alice went horse riding with a friend.

3. A: Why is Alice in the kitchen?
   B: Alice is eating her eggs and bread.
   A: Alice is eating her eggs and bread?
   B: Yes, Alice is eating her eggs and bread.

4. A: Does Alice often get reprimanded?
   B: Alice often gets reprimanded.
   A: Alice often gets reprimanded?
   B: Yes, Alice often gets reprimanded.

Block D (target sentences = 11 syllables long)

1. A: Who is Andrew?
   B: Andrew is an electrical engineer.
   A: Andrew is an electrical engineer?
   B: Yes, Andrew is an electrical engineer.

2. A: What is Andrew doing?
   B: Andrew is writing a letter to the bank.
   A: Andrew is writing a letter to the bank.
   B: Yes, Andrew is writing a letter to the bank.

3. A: Why is Andrew so quiet?
   B: Andrew feels relaxed after eating dinner.
   A: Andrew feels relaxed after eating dinner?
   B: Yes, Andrew feels relaxed after eating dinner.

4. A: Is Andrew a very bright entrepreneur?
   B: Andrew is a very bright entrepreneur.
   A: Andrew is a very bright entrepreneur?
   B: Yes, Andrew is a very bright entrepreneur.

Block E (target sentences = 13 syllables long)

1. A: Who is Morris?
   B: Morris is a member of the English Student Club.
   A: Morris is a member of the English Student Club?
   B: Yes, Morris is a member of the English Student Club.

2. A: What does Morris want to do?
   B: Morris wants to visit the old mansion on Monday.
   A: Morris wants to visit the old mansion on Monday?
   B: Yes, Morris wants to visit the old mansion on Monday.

3. A: Why is Morris so happy?
   B: Morris got a hundred percent on his English test.
   A: Morris got a hundred percent on his English test?
   B: Yes, Morris got a hundred percent on his English test.

4. A: Does Morris need to add olive oil to his rice noodles?
   B: Morris needs to add olive oil to his rice noodles.
   A: Morris needs to add olive oil to his rice noodles?
   B: Yes, Morris needs to add olive oil to his rice noodles.

A.2 Cantonese Stimuli

Block A: sentence initial (Tone 1 + Tone 6), sentence final (*si*), 5 syllables long

1. A: 汪義係邊個?
   Wong1 Ji6 hai6 bin1 go3?
   'Who is Wong Ji?'

   B: 汪義係老師。
   Wong1 Ji6 hai6 lou5 si1.
   'Wong Ji is a teacher.'

   A: 汪義係老師?
   Wong1 Ji6 hai6 lou5 si1?
   'Wong Ji is a teacher?'

   B: 係, 汪義係老師。
   Hai6, Wong1 Ji6 hai6 lou5 si1.
   'Yes, Wong Ji is a teacher.'

2. A: 汪義教乜嘢?
   Wong1 Ji6 gaau3 mat1 je5?
   'What does Wong Ji teach?'

   B: 汪義教歷史。
   Wong1 Ji6 gaau3 lik6 si2.
   'Wong Ji teaches history.'

   A: 汪義教歷史?
   Wong1 Ji6 gaau3 lik6 si2?
   'Wong Ji teaches history?'

   B: 係, 汪義教歷史。
   Hai6, Wong1 Ji6 gaau3 lik6 si2.
   'Yes, Wong Ji teaches history.'

3. A: 汪義點解重未來?
   Wong1 Ji6 dim2 gaai2 zung6 mei6 lei4?
   'Why hasn't Wong Ji come yet?'

   B: 汪義唔準時。
   Wong1 Ji6 m4 zeon2 si4.
   'Wong Ji is not on time.'

A: 汪義唔準時?
Wong1 Ji6 m4 zeon2 si4?
'Wong Ji is not on time?'

B: 係, 汪義唔準時。
Hai6, Wong1 Ji6 m4 zeon2 si4.
'Yes, Wong Ji is not on time.'

4. A: 汪義睇電視嗎?
Wong1 Ji6 tai2 din6 si6 maa1?
'Does Wong Ji watch TV?'

B: 汪義睇電視。
Wong1 Ji6 tai2 din6 si6.
'Wong Ji watches TV.'

A: 汪義睇電視?
Wong1 Ji6 tai2 din6 si6?
'Wong Ji watches TV?'

B: 係, 汪義睇電視。
Hai6, Wong1 Ji6 tai2 din6 si6.
'Yes, Wong Ji watches TV.'

Block B: sentence initial (Tone 4 + Tone 2), sentence final (*ji*), 7 syllables long

1. A: 余鎖係邊個?
Jyu4 So2 hai6 bin1 go3?
'Who is Jyu So?'

B: 余鎖係一個牙醫。
Jyu4 So2 hai6 jat1 go3 ngaa4 ji1.
'Jyu So is a dentist.'

A: 余鎖係一個牙醫?
Jyu4 So2 hai6 jat1 go3 ngaa4 ji1?
'Jyu So is a dentist?'

B: 係, 余鎖係一個牙醫。
Hai6, Jyu4 So2 hai6 jat1 go3 ngaa4 ji1.
'Yes, Jyu So is a dentist.'

2. A: 余鎖去買乜嘢?
Jyu4 So2 heoi3 maai5 mat1 je5?
'What is Jyu So going to buy?'

B: 余鎖去買按摩椅。
Jyu4 So2 heoi3 maai5 on3 mo1 ji2.
'Jyu So is going to buy a massage chair.'

A: 余鎖去買按摩椅?
Jyu4 So2 heoi3 maai5 on3 mo1 ji2?
'Jyu So is going to buy a massage chair?'

B: 係, 余鎖去買按摩椅。
Hai6, Jyu4 So2 heoi3 maai5 on3 mo1 ji2.
'Yes, Jyu So is going to buy a massage chair.'

3. A: 余鎖點解會識得佢?
Jyu4 So2 dim2 gaai2 wui5 sik1 dak1 keoi5?
'Why would Jyu So know him?'

B: 余鎖係佢嘅女兒。
Jyu4 So2 hai6 keoi5 ge3 neoi5 ji4.
'Jyu So is his daughter.'

A: 余鎖係佢嘅女兒?
Jyu4 So2 hai6 keoi5 ge3 neoi5 ji4?
'Jyu So is his daughter?'

B: 係, 余鎖係佢嘅女兒。
Hai6, Jyu4 So2 hai6 keoi5 ge3 neoi5 ji4.
'Yes, Jyu So is his daughter.'

4. A: 余鎖有一啲失意嗎?
Jyu4 So2 jau5 jat1 di1 sat1 ji3 maa1?
'Is Jyu So a bit disappointed?'

B: 余鎖有一啲失意。
Jyu4 So2 jau5 jat1 di1 sat1 ji3.
'Jyu So is a bit disappointed.'

A: 余鎖有一啲失意?
Jyu4 So2 jau5 jat1 di1 sat1 ji3?
'Jyu So is a bit disappointed?'

B: 係, 余鎖有一啲失意。
Hai6, Jyu4 So2 jau5 jat1 di1 sat1 ji3.
'Yes, Jyu So is a bit disappointed.'

Block C: sentence initial (Tone 6 + Tone 1), sentence final (*maa*), 9 syllables long

1. A: 路花係邊個?
Lou6 Faa1 hai6 bin1 go3?
'Who is Lou Faa?'

B: 路花係老朋友嘅姨媽。
Lou6 Faa1 hai6 lou5 pang4 jau5 ge3 ji4 maa1.
'Lou Faa is an old friend's aunt.'

A: 路花係老朋友嘅姨媽?
Lou6 Faa1 hai6 lou5 pang4 jau5 ge3 ji4 maa1?
'Lou Faa is an old friend's aunt?'

B: 係, 路花係老朋友嘅姨媽。
Hai6, Lou6 Faa1 hai6 lou5 pang4 jau5 ge3 ji4 maa1.
'Yes, Lou Faa is an old friend's aunt.'

2. A: 路花去做乜嘢?
Lou6 Faa1 heoi3 zou6 mat1 je5?
'What did Lou Faa go to do?

B: 路花跟咗朋友去騎馬。
Lou6 Faa1 gan1 zo2 pang4 jau5 heoi3 ke4 maa5.
'Lou Faa went horse riding with a friend.'

A: 路花跟咗朋友去騎馬?
Lou6 Faa1 gan1 zo2 pang4 jau5 heoi3 ke4 maa5?
'Lou Faa went horse riding with a friend?'

B: 係, 路花跟咗朋友去騎馬。
Hai6, Lou6 Faa1 gan1 zo2 pang4 jau5 heoi3 ke4 maa5.
'Yes, Lou Faa went horse riding with a friend.'

3. A: 路花點解喺廚房裡面?
Lou6 Faa1 dim2 gaai2 hai2 cyu4 fong2 leoi5 min6?
'Why is Lou Faa in the kitchen?'

B: 路花鍾意朝早食亞麻。
Lou6 Faa1 zung1 ji3 ziu1 zou2 sik6 aa3 maa4.
'Lou Faa likes to eat flax seed in the morning.'

A: 路花鍾意朝早食亞麻?
Lou6 Faa1 zung1 ji3 ziu1 zou2 sik6 aa3 maa4?
'Lou Faa likes to eat flax seed in the morning?'

B: 係, 路花鍾意朝早食亞麻。
Hai6, Lou6 Faa1 zung1 ji3 ziu1 zou2 sik6 aa3 maa4.
'Yes, Lou Faa likes to eat flax seed in the morning.'

4. A: 路花成日都被老闆罵嗎?
Lou6 Faa1 seng4 jat6 dou1 bei6 lou5 baan2 maa6 maa1?
'Does Lou Faa often get scolded by her boss?'

B: 路花成日都被老闆罵。
Lou6 Faa1 seng4 jat6 dou1 bei6 lou5 baan2 maa6.
'Lou Faa often gets scolded by her boss.'

A: 路花成日都被老闆罵?
Lou6 Faa1 seng4 jat6 dou1 bei6 lou5 baan2 maa6?
'Lou Faa often gets scolded by her boss?'

B: 係, 路花成日都被老闆罵。
Hai6, Lou6 Faa1 seng4 jat6 dou1 bei6 lou5 baan2 maa6.
'Yes, Lou Faa often gets scolded by her boss.'

Block D: sentence initial (Tone 2 + Tone 4), sentence final (*fu*(*k*)), 11 syllables long

1. A: 許狐係邊個?
Heoi2 Wu4 hai6 bin1 go3?
'Who is Heoi Wu?'

B: 許狐係一個好勤力嘅農夫。
Heoi2 Wu4 hai6 jat1 go3 hou2 kan4 lik6 ge3 nung4 fu1.
'Heoi Wu is a very hardworking farmer.'

A: 許狐係一個好勤力嘅農夫?
Heoi2 Wu4 hai6 jat1 go3 hou2 kan4 lik6 ge3 nung4 fu1?
'Heoi Wu is a very hardworking farmer?'

B: 係, 許狐係一個好勤力嘅農夫。
Hai6, Heoi2 Wu4 hai6 jat1 go3 hou2 kan4 lik6 ge3 nung4 fu1.
'Yes, Heoi Wu is a very hardworking farmer.'

2. A: 許狐做緊乜嘢?
Heoi2 Wu4 zou5 gan2 mat1 ye5?
'What is Heoi Wu doing?'

B: 許狐寫緊信畀加拿大政府。
Heoi2 Wu4 se2 gan2 seon3 bei2 gaa1 naa4 daai6 zing3 fu2.
'Heoi Wu is writing a letter to the Canadian government.'

A: 許狐寫緊信畀加拿大政府?
Heoi2 Wu4 se2 gan2 seon3 bei2 gaa1 naa4 daai6 zing3 fu2?
'Heoi Wu is writing a letter to the Canadian government?'

B: 係, 許狐寫緊信畀加拿大政府。
Hai6, Heoi2 Wu4 se2 gan2 seon3 bei2 gaa1 naa4 daai6 zing3 fu2.
'Yes, Heoi Wu is writing a letter to the Canadian government.'

3. A: 許狐點解咁靜?
Heoi2 Wu4 dim2 gaai2 gam3 zing6?
'Why is Heoi Wu so quiet?'

B: 許狐食咗飯覺得舒舒服服。
Heoi2 Wu4 sik6 zo2 fan4 gok3 dak1 syu1 syu1 fuk6 fuk6.
'Heoi Wu feels comfortable after eating rice.'

A: 許狐食咗飯覺得舒舒服服?
Heoi2 Wu4 sik6 zo2 fan4 gok3 dak1 syu1 syu1 fuk6 fuk6?
'Heoi Wu feels comfortable after eating rice?'

B: 係, 許狐食咗飯覺得舒舒服服。
Hai6, Heoi2 Wu4 sik6 zo2 fan4 gok3 dak1 syu1 syu1 fuk6 fuk6.
'Yes, Heoi Wu feels comfortable after eating rice.'

4. A: 許狐對自己嘅聰明好在乎嗎?
Heoi2 Wu4 deoi3 zi6 gei2 ge3 cung1 ming4 hou2 zoi6 fu4 maa1?
'Does Heoi Wu care about his own intelligence?'

B: 許狐對自己嘅聰明好在乎。
Heoi2 Wu4 deoi3 zi6 gei2 ge3 cung1 ming4 hou2 zoi6 fu4.
'Heoi Wu cares about his own intelligence.'

A: 許狐對自己嘅聰明好在乎?
Heoi2 Wu4 deoi3 zi6 gei2 ge3 cung1 ming4 hou2 zoi6 fu4?
'Heoi Wu cares about his own intelligence?'

B: 係, 許狐對自己嘅聰明好在乎。
Hai6, Heoi2 Wu4 deoi3 zi6 gei2 ge3 cung1 ming4 hou2 zoi6 fu4.
'Yes, Heoi Wu cares about his own intelligence.'

Block E: sentence initial (Tone 1 + Tone 1), sentence final (*fan*), 13 syllables long

1. A: 蘇仙係邊個?
Sou1 Sin1 hai6 bin1 go3?
'Who is Sou Sin?'

B: 蘇仙係愛民頓同學會嘅一部份。
Sou1 Sin1 hai6 oi3 man4 deon6 tung4 hok6 wui2 ge3 yat1 bou6 fan6.
'Sou Sin is part of The Edmonton Student Association.'

A: 蘇仙係愛民頓同學會嘅一部份?
Sou1 Sin1 hai6 oi3 man4 deon6 tung4 hok6 wui2 ge3 yat1 bou6 fan6?
'Sou Sin is part of The Edmonton Student Association?'

B: 係, 蘇仙係愛民頓同學會嘅一部份。
Hai6, Sou1 Sin1 hai6 oi3 man4 deon6 tung4 hok6 wui2 ge3 yat1 bou6 fan6.
'Yes, Sou Sin is part of The Edmonton Student Association.'

2. A: 蘇仙想做乜嘢?
Sou1 Sin1 seong2 zou5 mat1 ye5?
'What does Sou Sin want to do?'

B: 蘇仙想同佢妹妹星期六去上墳。
Sou1 Sin1 seong2 tung4 keoi5 mui4 mui2 sing1 kei4 luk6 heoi3 soeng5 fan4.
'Sou Sin wants to go visit her ancestor's grave with her sister on Saturday.'

A: 蘇仙想同佢妹妹星期六去上墳?
Sou1 Sin1 seong2 tung4 keoi5 mui4 mui2 sing1 kei4 luk6 heoi3 soeng5 fan4?
'Sou Sin wants to go visit her ancestor's grave with her sister on Saturday?'

B: 係, 蘇仙想同佢妹妹星期六去上墳。
Hai6, Sou1 Sin1 seong2 tung4 keoi5 mui4 mui2 sing1 kei4 luk6 heoi3 soeng5 fan4.
'Yes, Sou Sin wants to go visit her ancestor's grave with her sister on Saturday.'

3. A: 蘇仙點解特別咁開心?
Sou1 Sin1 dim2 gaai2 dak6 bit6 gam3 hoi1 sam1?
'Why is Sou Sin so happy?'

B: 蘇仙今日英文考試得到一百分。
Sou1 Sin1 gam1 jat6 jing1 man2 haau2 si5 dak1 dou2 jat1 baak3 fan1.
'Sou Sin got a hundred percent on her English test today.'

A: 蘇仙今日英文考試得到一百分?
Sou1 Sin1 gam1 jat6 jing1 man2 haau2 si5 dak1 dou2 jat1 baak3 fan1?
'Sou Sin got a hundred percent on her English test today?'

B: 係, 蘇仙今日英文考試得到一百分。
Hai6, Sou1 Sin1 gam1 jat6 jing1 man2 haau2 si5 dak1 dou2 jat1 baak3 fan1.
'Yes, Sou Sin got a hundred percent on her English test today.'

4. A: 蘇仙需要喺米粉上面加辣椒粉嗎?
Sou1 Sin1 seoi1 jiu3 hai2 mai5 fan2 soeng6 min6 gaa1 laat6 ziu1 fan2 maa1?
'Does Sou Sin need to add chili powder on top of the rice noodles?'

B: 蘇仙需要喺米粉上面加辣椒粉。
Sou1 Sin1 seoi1 jiu3 hai2 mai5 fan2 soeng6 min6 gaa1 laat6 ziu1 fan2.
'Sou Sin needs to add chili powder on top of the rice noodles.'

A: 蘇仙需要喺米粉上面加辣椒粉?
Sou1 Sin1 seoi1 jiu3 hai2 mai5 fan2 soeng6 min6 gaa1 laat6 ziu1 fan2?
'Sou Sin needs to add chili powder on top of the rice noodles?'

B: 係, 蘇仙需要喺米粉上面加辣椒粉。
Hai6, Sou1 Sin1 seoi1 jiu3 hai2 mai5 fan2 soeng6 min6 gaa1 laat6 ziu1 fan2.
'Yes, Sou Sin needs to add chili powder on top of the rice noodles.'

A.3 Mandarin Stimuli

Block A: sentence initial (Tone 1 + Tone 3), sentence final (*shi*), 5 syllables long

1. A: 汪五是谁?
   Wang1 Wu3 shi4 shei2?
   'Who is Wang Wu?'

   B: 汪五是老师。
   Wang1 Wu3 shi4 lao3 shi1.
   'Wang Wu is a teacher.'

   A: 汪五是老师?
   Wang1 Wu3 shi4 lao3 shi1?
   'Wang Wu is a teacher?'

   B: 是, 汪五是老师。
   Shi4, Wang1 Wu3 shi4 lao3 shi1.
   'Yes, Wang Wu is a teacher.'

2. A: 汪五教什么?
   Wang1 Wu3 jiao4 shen2 me?
   'What does Wang Wu teach?'

   B: 汪五教历史。
   Wang1 Wu3 jiao4 li4 shi3.
   'Wang Wu teaches history.'

   A: 汪五教历史?
   Wang1 Wu3 jiao4 li4 shi3?
   'Wang Wu teaches history?'

   B: 是, 汪五教历史。
   Shi4, Wang1 Wu3 jiao4 li4 shi3.
   'Yes, Wang Wu teaches history.'

3. A: 汪五为什么还没来?
   Wang1 Wu3 wei4 shen2 me hai2 mei2 lai2?
   'Why hasn't Wang Wu come yet?'

   B: 汪五不准时。
   Wang1 Wu3 bu4 zhun3 shi2.
   'Wang Wu is not on time.'

A: 汪五不准时?
Wang1 Wu3 bu4 zhun3 shi2?
'Wang Wu is not on time?'

B: 是, 汪五不准时。
Shi4, Wang1 Wu3 bu4 zhun3 shi2.
'Yes, Wang Wu is not on time.'

4. A: 汪五看电视吗?
Wang1 Wu3 kan4 dian4 shi4 ma?
'Does Wang Wu watch TV?

B: 汪五看电视。
Wang1 Wu3 kan4 dian4 shi4.
'Wang Wu watches TV.'

A: 汪五看电视?
Wang1 Wu3 kan4 dian4 shi4?
'Wang Wu watches TV?'

B: 是, 汪五看电视。
Shi4, Wang1 Wu3 kan4 dian4 shi4.
'Yes, Wang Wu watches TV.'

Block B: sentence initial (Tone 4 + Tone 2), sentence final (*yi*), 7 syllables long

1. A: 叶十是谁?
Ye4 Shi2 shi4 shei2?
'Who is Ye Shi?'

B: 叶十是一个牙医。
Ye4 Shi2 shi4 yi1 ge4 ya2 yi1.
'Ye Shi is a dentist.'

A: 叶十是一个牙医?
Ye4 Shi2 shi4 yi1 ge4 ya2 yi1?
'Ye Shi is a dentist?'

B: 是, 叶十是一个牙医。
Shi4, Ye4 Shi2 shi4 yi1 ge4 ya2 yi1.
'Yes, Ye Shi is a dentist.'

2. A: 叶十去买什么?
Ye4 Shi2 qu4 mai3 shen2 me?
'What is Ye Shi going to buy?'

B: 叶十去买按摩椅。
Ye4 Shi2 qu4 mai3 an4 mo2 yi3.
'Ye Shi is going to buy a massage chair.'

A: 叶十去买按摩椅?
Ye4 Shi2 qu4 mai3 an4 mo2 yi3?
'Ye Shi is going to buy a massage chair?'

B: 是, 叶十去买按摩椅。
Shi4, Ye4 Shi2 qu4 mai3 an4 mo2 yi3.
'Yes, Ye Shi is going to buy a massage chair.'

3. A: 叶十为什么会认识他?
Ye4 Shi2 wei4 shen2 me hui4 ren4 shi ta1?
'Why would Ye Shi know him?'

B: 叶十是他的阿姨。
Ye4 Shi2 shi4 ta1 de a1 yi2.
'Ye Shi is his aunt.'

A: 叶十是他的阿姨?
Ye4 Shi2 shi4 ta1 de a1 yi2?
"Ye Shi is his aunt?'

B: 是, 叶十是他的阿姨。
Shi4, Ye4 Shi2 shi4 ta1 de a1 yi2.
'Yes, Ye Shi is his aunt.'

4. A: 叶十有一点失忆吗?
Ye4 Shi2 you3 yi1 dian3 shi1 yi4 ma?
'Is Ye Shi slightly forgetful?'

B: 叶十有一点失忆。
Ye4 Shi2 you3 yi1 dian3 shi1 yi4.
'Ye Shi is slightly forgetful.'

A: 叶十有一点失忆?
Ye4 Shi2 you3 yi1 dian3 shi1 yi4?
'Ye Shi is slightly forgetful?'

B: 是, 叶十有一点失忆。
Shi4, Ye4 Shi2 you3 yi1 dian3 shi1 yi4.
'Yes, Ye Shi is slightly forgetful.'

Block C: sentence initial (Tone 3 + Tone 1), sentence final (*ma*), 9 syllables long

1. A: 李一是谁?
Li3 Yi1 shi4 shei2?
'Who is Li Yi?'

B: 李一是老朋友的姨妈。
Li3 Yi1 shi4 lao3 peng2 you de yi2 ma1.
'Li Yi is an old friend's aunt.'

A: 李一是老朋友的姨妈?
Li3 Yi1 shi4 lao3 peng2 you de yi2 ma1?
'Li Yi is an old friend's aunt?'

B: 是, 李一是老朋友的姨妈。
Shi4, Li3 Yi1 shi4 lao3 peng2 you de yi2 ma1.
'Yes, Li Yi is an old friend's aunt.'

2. A: 李一去做什么?
Li3 Yi1 qu4 zuo4 shen2 me?
'What did Li Yi go to do?'

B: 李一跟了朋友去骑马。
Li3 Yi1 gen1 le peng2 you qu4 qi2 ma3.
'Li Yi went horse riding with a friend.'

A: 李一跟了朋友去騎马?
Li3 Yi1 gen1 le peng2 you qu4 qi2 ma3?
'Li Yi went horse riding with a friend?'

B: 是, 李一跟了朋友去騎马。
Shi4, Li3 Yi1 gen1 le peng2 you qu4 qi2 ma3.
'Yes, Li Yi went horse riding with a friend.'

3. A: 李一为什么在厨房里面?
Li3 Yi1 wei4 shen2 me zai4 chu2 fang2 li3 mian4?
'Why is Li Yi in the kitchen?'

B: 李一喜欢早上吃亚麻。
Li3 Yi1 xi3 huan zao3 shang chi1 ya4 ma2.
'Li Yi likes to eat flax seed in the morning.'

A: 李一喜欢早上吃亚麻?
Li3 Yi1 xi3 huan zao3 shang chi1 ya4 ma2?
'Li Yi likes to eat flax seed in the morning?'

B: 是, 李一喜欢早上吃亚麻。
Shi4, Li3 Yi1 xi3 huan zao3 shang chi1 ya4 ma2.
'Yes, Li Yi likes to eat flax seed in the morning.'

4. A: 李一动不动被上司骂吗?
Li3 Yi1 dong4 bu dong4 bei4 shang4 si ma4 ma?
'Does Li Yi often get scolded by her boss?'

B: 李一动不动被上司骂。
Li3 Yi1 dong4 bu dong4 bei4 shang4 si ma4.
'Li Yi often gets scolded by her boss.'

A: 李一动不动被上司骂?
Li3 Yi1 dong4 bu dong4 bei4 shang4 si ma4?
'Li Yi often gets scolded by her boss?'

B: 是, 李一动不动被上司骂。
Shi4, Li3 Yi1 dong4 bu dong4 bei4 shang4 si ma4.
'Yes, Li Yi often gets scolded by her boss.'

Block D: sentence initial (Tone 2 + Tone 4), sentence final (*fu*), 11 syllables long

1. A: 吴二是谁?
Wu2 Er4 shi4 shei2?
'Who is Wu Er?'

B: 吴二是一个很努力的农夫。
Wu2 Er4 shi4 yi1 ge4 hen3 nu3 li4 de nong2 fu1.
'Wu Er is a very hardworking farmer.'

A: 吴二是一个很努力的农夫?
Wu2 Er4 shi4 yi1 ge4 hen3 nu3 li4 de nong2 fu1?
'Wu Er is a very hardworking farmer?'

B: 是, 吴二是一个很努力的农夫。
Shi4, Wu2 Er4 shi4 yi1 ge4 hen3 nu3 li4 de nong2 fu1.
'Yes, Wu Er is a very hardworking farmer.'

2. A: 吴二在做什么?
Wu2 Er4 zai4 zuo4 shen2 me?
'What is Wu Er doing?'

B: 吴二在写信给加拿大政府。
Wu2 Er4 zai4 xie3 xin4 gei3 jia1 na2 da4 zheng4 fu3.
'Wu Er is writing a letter to the Canadian government.'

A: 吴二在写信给加拿大政府?
Wu2 Er4 zai4 xie3 xin4 gei3 jia1 na2 da4 zheng4 fu3?
'Wu Er is writing a letter to the Canadian government?'

B: 是, 吴二在写信给加拿大政府。
Shi4, Wu2 Er4 zai4 xie3 xin4 gei3 jia1 na2 da4 zheng4 fu3.
'Yes, Wu Er is writing a letter to the Canadian government.'

3. A: 吴二为什么这样安静?
Wu2 Er4 wei3 shen2 me zhe4 yang4 an1 jing4?
'Why is Wu Er so quiet?'

B: 吴二吃了饭觉得舒舒服服。
Wu2 Er4 chi1 le fan4 jue2 de2 shu1 shu1 fu2 fu2.
'Wu Er feels comfortable after eating rice.'

A: 吴二吃了饭觉得舒舒服服?
Wu2 Er4 chi1 le fan4 jue2 de2 shu1 shu1 fu2 fu2?
'Wu Er feels comfortable after eating rice?'

B: 是, 吴二吃了饭觉得舒舒服服。
Shi4, Wu2 Er4 chi1 le fan4 jue2 de2 shu1 shu1 fu2 fu2.
'Yes, Wu Er feels comfortable after eating rice.'

4. A: 吴二是一个很聪明的师傅吗?
Wu2 Er4 shi4 yi1 ge4 hen3 cong1 ming2 de shi1 fu4 ma?
'Is Wu Er a very smart master?'

B: 吴二是一个很聪明的师傅。
Wu2 Er4 shi4 yi1 ge4 hen3 cong1 ming2 de shi1 fu4.
'Wu Er is a very smart master.'

A: 吴二是一个很聪明的师傅?
Wu2 Er4 shi4 yi1 ge4 hen3 cong1 ming2 de shi1 fu4?
'Wu Er is a very smart master?'

B: 是, 吴二是一个很聪明的师傅。
Shi4, Wu2 Er4 shi4 yi1 ge4 hen3 cong1 ming2 de shi1 fu4.
'Yes, Wu Er is a very smart master.'

Block E: sentence initial (Tone 1 + Tone 1), sentence final (*fen*), 13 syllables long

1. A: 苏三是谁?
Su1 San1 shi4 shei2?
'Who is Su San?'

B: 苏三是爱民顿同学会的一部份。
Su1 San1 shi4 Ai4 Min2 Dun4 tong2 xue2 hui4 de yi1 bu4 fen4.
'Su San is part of the Edmonton Student Association.'

A: 苏三是爱民顿同学会的一部份?
Su1 San1 shi4 Ai4 Min2 Dun4 tong2 xue2 hui4 de yi1 bu4 fen4?
'Su San is part of the Edmonton Student Association?'

B: 是, 苏三是爱民顿同学会的一部份。
Shi4, Su1 San1 shi4 Ai4 Min2 Dun4 tong2 xue2 hui4 de yi1 bu4 fen4.
'Yes, Su San is part of the Edmonton Student Association.'

2. A: 苏三想做什么?
Su1 San1 xiang3 zuo4 shen2 me?
'What does Su San want to do?'

B: 苏三想和她妹妹星期六去上坟。
Su1 San1 xiang3 he2 ta1 mei4 mei xing1 qi1 liu4 qu4 shang4 fen2.
'Su San wants to go visit her ancestor's grave with her sister on Saturday.'

A: 苏三想和她妹妹星期六去上坟?
Su1 San1 xiang3 he2 ta1 mei4 mei xing1 qi1 liu4 qu4 shang4 fen2?
'Su San wants to go visit her ancestor's grave with her sister on Saturday?'

B: 是, 苏三想和她妹妹星期六去上坟。
Shi4, Su1 San1 xiang3 he2 ta1 mei4 mei xing1 qi1 liu4 qu4 shang4 fen2.
'Yes, Su San wants to go visit her ancestor's grave with her sister on Saturday.'

3. A: 苏三为什么特别高兴?
Su1 San1 wei4 shen2 me te4 bie2 gao1 xing4?
'Why is Su San so happy?'

B: 苏三今天英文考试得到一百分。
Su San1 jin1 tian1 ying1 wen2 kao3 shi4 de2 dao4 yi1 bai3 fen1.
'Su San got a hundred percent on her English test today.'

A: 苏三今天英文考试得到一百分?
Su San1 jin1 tian1 ying1 wen2 kao3 shi4 de2 dao4 yi1 bai3 fen1?
'Su San got a hundred percent on her English test today?'

B: 是, 苏三今天英文考试得到一百分。
Shi4, Su1 San1 jin1 tian1 ying1 wen2 kao3 shi4 de2 dao4 yi1 bai3 fen1.
'Yes, Su San got a hundred percent on her English test today.'

4. A: 苏三需要在米粉上面加辣椒粉吗?
Su1 San1 xu1 yao4 zai4 mi3 fen3 shang4 mian4 jia1 la4 jiao1 fen3 ma?
'Does Su San need to add chili powder on top of the rice noodles?'

B: 苏三需要在米粉上面加辣椒粉。
Su1 San1 xu1 yao4 zai4 mi3 fen3 shang4 mian4 jia1 la4 jiao1 fen3.
'Su San needs to add chili powder on top of the rice noodles.'

A: 苏三需要在米粉上面加辣椒粉?
Su1 San1 xu1 yao4 zai4 mi3 fen3 shang4 mian4 jia1 la4 jiao1 fen3?
'Su San needs to add chili powder on top of the rice noodles?'

B: 是, 苏三需要在米粉上面加辣椒粉。
Shi4, Su1 San1 xu1 yao4 zai4 mi3 fen3 shang4 mian4 jia1 la4 jiao1 fen3.
'Yes, Su San needs to add chili powder on top of the rice noodles.'

Appendix B: Background Questionnaire

Project: An exemplar-based model of intonation perception Participant # _____
Researcher: Una Chow
Supervisor: Dr. Stephen Winters

Research Study Participation Background Questionnaire

1. Age: _____ 2. Gender: _____

3. Where have you lived, and how old were you when you lived there?
(specify country and region/city)

4. Where are your parents from?

5. What language did your parents speak while you were growing up?

6. List any languages that you might know (including your native language), the age at which you first started learning that language (0 = from birth), and your speaking, listening, and reading proficiency of that language (where 1 = poor and 5 = excellent).

Language	Speaking	Listening	Reading	Age first learned
_____	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	_____
_____	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	_____
_____	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	_____
_____	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	_____
_____	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	_____

7. What language(s) do you currently speak on a daily basis?

8. Do you have any speech, hearing, or visual impairments? If so, please describe.

9. Do you play any musical instruments? If so, which ones?

Appendix C: Letter of Copyright Permission

April 27, 2017

To Whom It May Concern:

As Una Chow's co-author on a number of papers whose content forms parts of this thesis, I hereby give permission to her to submit the thesis to Library and Archives Canada. I also acknowledge that Una has informed me that Library and Archives Canada may reproduce and make available this thesis to the public for non-commercial purposes.

The papers which form part of this thesis that we have collaborated on include the following:

Chow, U. Y., & Winters, S. J. (2015). Exemplar-based classification of statements and questions in Cantonese. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. Paper number 0987.1-5.

Chow, U. Y., & Winters, S. J. (2016). Perception of intonation in Cantonese: Native listeners versus an exemplar-based model. *Proceedings of the 2016 Annual Conference of the Canadian Linguistic Association*.

Chow, U. Y., & Winters, S. J. (2016). Perception of statement and question intonation: Cantonese versus Mandarin. *Proceedings of the 16th Australasian International Conference on Speech Science and Technology*, 13-16.

Chow, U. Y., & Winters, S. J. (2016). The role of the final tone in signaling statements and questions in Mandarin. *Proceedings of the 5th International Symposium on Tonal Aspects of Languages*, 167-171.

Sincerely,

Stephen J. Winters